ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

PRODUCTION NOTE

University of Illinois at
Urbana-Champaign Library
Large-scale Digitization Project, 2007.

Technical Report No. 253

THE NUMBER OF WORDS IN PRINTED SCHOOL ENGLISH

William E. Nagy and Richard C. Anderson
University of Illinois at Urbana-Champaign

July 1982

Center for the Study of Reading

TECHNICAL REPORTS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
51 Gerty Drive
Champaign, Illinois 61820

BOLT BERANEK AND NEWMAN INC.
50 Moulton Street
Cambridge, Massachusetts 02238

The National Institute of Education
U.S. Department of Education
Washington, D.C. 20208

CENTER FOR THE STUDY OF READING

Technical Report No. 253

THE NUMBER OF WORDS IN PRINTED SCHOOL ENGLISH

William E. Nagy and Richard C. Anderson
University of Illinois at Urbana-Champaign

July 1982

University of Illinois
at Urbana-Champaign
51 Gerty Drive
Champaign, Illinois 61820

Bolt Beranek and Newman Inc.
50 Moulton Street
Cambridge, Massachusetts 02238

The research reported herein was supported by the National Institute of Education under Contract No. US-NIE-C-400-76-0116.

Abstract

The purpose of this research was to determine the number of distinct words in printed school English. A detailed analysis was done of a 7,260 word sample from the Carroll, Davies and Richman, Word Frequency Book. Projecting from the sample to the total vocabulary of school English, our best estimate is that it contains about 88,500 distinct words. Furthermore, for every word a child learns, we estimate that there are an average of one to three additional related words that should also be understandable to the child, the exact number depending on how well the child is able to utilize context and morphology to induce meanings. Based on our analysis, a reconciliation of estimates of children's vocabulary size was undertaken, which showed that the extreme divergence in estimates is due mainly to the definition of "word" adopted. Our findings indicate that even the most ruthlessly systematic direct vocabulary instruction could neither account for a significant proportion of all the words children actually learn, nor cover more than a modest proportion of the words they will encounter in school reading materials.

The Number of Words in Printed School English

Determining the absolute size of individuals' vocabularies is of more than purely theoretical interest. If a student must learn 8,000 words by his or her senior year in high school, this goal might be reached via an ambitious program of direct instruction. If, on the other hand, the number of words to be learned were closer to 80,000, this goal would be beyond the reach of even the most intensive direct instruction that could be accomplished in the time available. The absolute size of vocabularies also has implications for theories of learning and language acquisition. If some seventh graders have vocabularies of over 50,000 words, as is estimated by some researchers, a theory of language acquisition must include mechanisms that could account for this phenomenal accomplishment.

There is in fact a substantial lack of agreement among researchers as to the absolute size of vocabulary at any given age or level of development (see Anderson & Freebody, 1981). For example, estimates of average total vocabulary size at third grade range from 2,000 words (Dupuy, 1974) to 25,000 words (M. K. Smith, 1941). The same two researchers estimate the vocabularies of seventh graders to be around 4,760 and 51,000 words, respectively. Some of the reasons for such large disparities between estimates are the source of words (e.g., what dictionary or corpus to take as representing English vocabulary, and how to choose a representative sample), testing methods (disagreements about

when a word can be counted as "known," and how to test such knowledge), and the definition of "word" adopted (disagreements about, for example, whether to include proper names, or under what conditions to count derived words as separate items).

It is with the third of these issues that we will primarily be concerned here. Our goal is to answer the question "How many different words are there?" in a number of ways, for a variety of criteria for defining "distinct words." This will allow us to reconcile estimates of vocabulary size based on different criteria for counting words. Our technique will be to recalibrate previous estimates using benchmarks derived from a corpus that we have analyzed in depth.

A Corpus of Words Representative of Printed School English

Dictionaries are often used as a starting point for building tests to estimate vocabulary size, although, as Carroll (1964) pointed out, this is a questionable practice. The organization and inclusion or exclusion of items in a dictionary will reflect not only linguistic principles, but also diverse practical demands such as page format and limitations on overall size. And the estimates of vocabulary size that a given test produces are related to the size of the dictionary that was used in constructing the test (Lorge & Chall, 1963; Hartman, 1941). It should be apparent that a dictionary is an unstable base from which to estimate vocabulary size.

Further variation is introduced in the selection of items from the dictionary. Researchers differ in whether categories such as proper names, technical terms, or scientific names of flora and fauna are included, and in the criteria for determining which derived words are to be counted as separate items.

Constructing or evaluating a test which attempts to measure absolute vocabulary size, therefore, depends on the answer to three questions: What source of words should be used, what types of words should be included or excluded, and under what conditions related words should be grouped together or treated as separate items. In this paper we will attempt to give principled answers to these questions. The goal is estimates of vocabulary size that are interpretable in terms of their implications for vocabulary instruction.

We have chosen as our source of words Carroll, Davies, and Richman's (1971) American Heritage Word Frequency Book (henceforth, the WFB). This book is based on the American Heritage Intermediate Corpus, which contains 5,088,721 words of running text from over a thousand items of published materials in use in schools. These were selected on the basis of a careful survey "to represent, as nearly as possible, the range of required and recommended reading to which students are exposed in school grades three through nine in the United States" (p. xxi). The materials sampled included textbooks, workbooks, kits, novels, poetry, general nonfiction, encyclopedias, and magazines. The WFB summarizes the largest and most recent corpus of the written language children encounter in

school. Furthermore, Carroll, Davies, and Richman have been able to use the corpus to determine properties not just of the vocabulary contained in the WFB, but of the total vocabulary of the type of materials from which the sample was collected. This total vocabulary is a theoretical construct, but its overall size (and several other properties) can be predicted with a substantial degree of confidence. Thus, our analysis can be generalized not just to the vocabulary in the WFB, but to the entire population of which the WFB constitutes a representative sample. Because of the way that the American Heritage Intermediate Corpus was collected, we can justifiably refer to this population as "printed school English" (with the restriction to grades three through nine understood).

"Printed school English," in this sense, gives us the basis for an operational definition of the total vocabulary of English, keeping in mind that we are restricting ourselves to written language intended largely for children. A vocabulary test based on this material could not be taken as a measure of a child's oral vocabulary, but would certainly be appropriate as a measure of a child's reading vocabulary.

One might be concerned at this point that written language intended for children is too restricted in vocabulary. This concern seems reasonable, but as it turns out it is not warrented. As we will see, even an unabridged dictionary gives a more limited picture of English vocabulary than do the projections of Carroll and his associates from their sample to the total vocabulary of written materials used in schools.

On Defining the Concept "Word"

Absolute vocabulary size can only be discussed in terms of some theory of relatedness among words. For example, the WFB is described as containing 86,741 different words, or types. However, since the corpus was sorted by computer, "word" is defined as a graphically distinct sequence of characters bounded right and left by a space. By this definition, doctor, Doctor, and DOCTOR are counted as three different words. Obviously, a psychologically more realistic definition of "word" will count these three types as instances of the "same word."

Dictionaries have traditionally treated regular inflectional varients, for example, walk, walks, and walked, as forms of the same word. This is pedagogically justifiable; by the time children reach first grade, they have normally learned the basics of English inflection. If a child has learned the word antelope, no separate instruction about the plural antelopes is needed; children can automatically apply the rules of regular pluralization to new forms (Berko, 1958).

Some dictionaries take other types of relatedness into account when grouping words into entries. Many list semantically transparent derivatives as subentries. For example, the American Heritage School Dictionary gives meekness and meekly as subentries under meek without further definition. Along similar lines, Thorndike (1921) grouped adverbs ending in -ly under their base forms, thus counting sadly and sad as one word. From a theoretical perspective, Aronoff (1976) argued that words derived by totally productive word formation processes (e.g., -ness, -ly) should not be given separate entries in the lexicon.

However, there is a great variety of types and degrees of relatedness among words that might be taken into consideration when estimating vocabulary size, ranging from the transparent cases just mentioned to more obscure relationships such as that between quiet and acquiesce. And there has been little agreement among vocabulary researchers as to how different types of relatedness among words should be treated. The extremes run from counting inflectional varients as separate words on the one hand, to a radical grouping such as in Dupuy (1974), who excluded from his count of "Basic Words" almost all suffixed, prefixed, and compound items, since these could in some sense be considered to be derived from more basic words, and hence at least partially redundant. It should be clear that decisions concerning how words should be counted will be a major factor in determining the magnitude of estimates of vocabulary size.

Previous analyses of relatedness among words have not provided an adequate basis for meaningful measures of absolute vocabulary size; they each suffer from at least one of a number of weaknesses. Many take an etymological or historical, rather than synchronic, approach to relationships among words, positing relationships based on information not available to the normal language learner. Some statistical analysis of word formation have been limited to prefixes, or to suffixes, or perhaps both of these, while neglecting compounding. Previous studies have usually adopted a single criterion of relatedness among words, without distinguishing types or degrees of relatedness. Some studies are

based on wordlists such as Thorndike and Lorge (1944) which are now outdated.

Becker, Dixon and Anderson-Inman (1980) have perhaps come closest to our purposes in their analysis of a vocabulary list derived by modifying and updating Thorndike and Lorge (1944). They have analysed a list of 25,782 words into morphographs (minimal "meaningful" units of written English), and assigned each word a root word which represents the smallest word from which a given word can be "semantically derived." This root word analysis does define patterns of interrelatedness among words to a certain extent. For example, divide, divided, dividend, dividers, dividing, divisible, division, divisional, and divisor are related in that all have been assigned the same root word divide.

However, in their analysis, there are no distinctions made between possible types or degrees of relatedness. Also, relatedness is defined on an etymological rather than synchronic basis. For example, millenium was assigned the root word annual. It is certainly possible for a historical linguist to see the relationship in form between these two words, but dubious that the normal speaker of English, armed only with such knowledge of morphology as can be gained from words currently in the language, would find any but a semantic relationship. Animism and animosity were assigned the root word anima; in this case, the relationship in form may be obvious, but the semantic relationship is rather distant. In the case of polynomial and its root word name, both the formal and semantic relationships are tenuous.

Analyses of affixes, for example, Thorndike (1941) or Stauffer (1942), have also typically been done on an etymological basis, e.g., segmenting fragile into a root frag- and the suffix -ile, or deceive into the prefix de- and the root -ceive. An exception to this is found in Harwood and Wright (1956) who specify in their counts which suffixed forms have a free base (e.g. acceptable) and which do not (e.g. amiable). However, while these analyses do give an indication of the extent to which some suffixes account for a portion of the overall vocbulary, they do not provide a basis for estimating the overall size of vocabulary, that is, they do not tell us what percentage of words actually are derivable using a given suffix.

Rhode and Cronnell (1977) have analysed a set of vocabulary items especially compiled to cover words used in grades K-6. However, their analysis, while including much useful information, focuses on types of letter-sound correspondences, so that their definitions of "prefix" and "suffix" are not in terms of productive word-formation processes in today's English. For example, their list of suffixes includes the om of bottom and the il of peril.

In our analyses, we will approach the question of relatedness among words not solely in terms of similarity of form, or in terms of etymological relationships, but rather, in terms of the relative ease or difficulty with which a child could either learn the meaning of that word, or infer its meaning in context while reading. Also, we will define different types and degrees of relatedness among words, so that we

can adjust our definitions of "related" and "distinct" to match the knowledge of word-relatedness of children at a given age or ability level.

Method

The data and statistical analyses in the WFB provide a reliable starting point for investigating the vocabulary of printed school English. However, the definition of "word" adopted for the purpose of compiling the WFB is, as the authors would freely admit, inappropriate for any linguistic or pedagogical estimate of vocabulary size. Our goal, then, is to categorize the different types of words in the WFB, and how they are related to each other, in order to arrive at a meaningful estimate of the number of different words in printed school English.

A random sample of 7,260 words was drawn from the 86,741 words in the WFB. This sample consists of 121 chunks of 60 contiguous words. The chunks were approximately evenly distributed throughout the alphabetical list. Contiguous groups of words were taken because related words are usually (but not always) close to each other in an alphabetical listing.

Table 1 gives an example of a group of related words, or "word family," that is found in one of the chunks in our sample. The pattern of interrelationships among these items is somewhat complex. It might be represented graphically as in Figure 1. This figure shows that there are multiple-branching structures, and that two words may be related via one or more intervening words. This figure does not distinguish between different types or degrees of relatedness among words. A more complete

representation would specify, for example, that the relationship between add and ADD is one of capitalization, while the relationship between addition and additional is suffixation.

Insert Table 1 and Figure 1 about here.

The set of possible relationships can be represented in terms of pairs of words, each pair representing two words which are adjacent and connected by a line in Figure 1. This type of representation, as depicted in Table 2, was used in our analyses. For each word in our sample, its "immediate ancestor" was found, that is, the word to which it is most closely related and which is in some sense more basic than the target word.

Insert Table 2 about here.

In the majority of cases, the identity of the immediate ancestor is not problematic. For an inflected form, e.g., adds, the immediate ancestor is the uninflected stem or infinitive, add. For the past tense, it would be the present (infinitive) form as well. For plurals, the immediate ancestor is the singular. For forms with a prefix, e.g., unknown, the immediate ancestor is the unprefixed form, known. For forms with a suffix, additional, the immediate ancestor is the form without the suffix, addition. For compounds, e.g., addition-subtraction, there are

two immediate ancestors, one for each part, in this case, addition and subtraction.

More problematic cases were treated as follows: If a word has both a prefix and suffix, as does undecided, one choses as the immediate ancestor the form that is semantically closest. In this case, there is no word *undecide, so that only one analysis is possible: undecided has as its immediate ancestor decided, which in turn has as its immediate ancestor decide. In a case such as reactivation there are two reasonable analyses. On the other hand, both analyses arrive at activate as an ancestor, and the choice will not make any difference in terms of the ultimate count of prefixes and suffixes.

In some relationships, for example, that between multiple and the verb multiply, it is difficult to say which item is more "basic" than the other. We recognize all the dangers and complications of saying that one word is "derived from" another. For the purposes of analysing patterns of interrelatedness among the words in the corpus, it is necessary to break down the relationships into assymetrical dyads; however, we assign no theoretical weight to the directionality of the relationship.

In some cases, the immediate ancestor of a given item was not found in the corpus. For example, abatement and abates are both found, but not abate. In this case, the item abate was added to the list, and flagged as a "missing ancestor." Sometimes intermediate forms were missing. In the group of words in Tables 1 and 2, for example, if the word addend had not occurred in the corpus, the relationship between addends and add

would have involved two steps, suffixation and pluralization. In our analyses we supplied such "missing links" wherever necessary, flagging them to mark that they were not in the original list of words from the WFB.

For each pair of items, the relationship between them was categorized. The basic categories used in our analyses are listed and exemplified in Table 3. A more detailed description of these categories and their special subcategories is found in Appendix A.

Insert Table 3 about here.

Coding Semantic Relatedness

In addition to distinguishing among different types of formal relationships between a word and its immediate ancestor (e.g., suffixation, prefixation, compounding), our coding system categorizes the semantic relationship between the two. For some pairs, e.g., tranquil/tranquility, the semantic relationship is fairly direct. For other pairs of words, it is more distant, e.g., fun/funny, live/lively, or descend/condescend.

An immediate problem in trying to characterize the semantic relationship between two words is the fact that one or both of them may have a number of meanings. Before one can describe the semantic relationship between the two, one must first decide which two meanings are to be compared.

We have tackled this problem in our coding system by representing the semantic relationship between two words in terms of two dimensions. The first represents the semantic relationship between the two most similar meanings of the two words. The second represents the relationship between the two most similar familiar meanings of the two words.

What constitutes a "familiar" meaning was necessarily defined in a rather impressionistic fashion. Basically, a "familiar" meaning was defined as one which would be likely to occur to an individual when seeing the word out of context. Given that people are relatively accurate at intuitively assessing the relative frequencies of different words (cf. Carroll, 1971, and Carroll et. al., 1971) it was hoped that an intuitive judgement as to the relative frequencies of word meanings would be adequate for the distinctions which were necessary to make here.

The words carry and carriage illustrate well the distinction we have made between the relationship of the two most similar meanings and the relationship of the two most similar familiar meanings. The two most similar meanings of these words might be the following:

carry: to hold or move (the body or part of the body) in a certain way

carriage: the manner in which the body is held; posture

These definitions are from the American Heritage School Dictionary, which is based on the American Heritage Intermediate Corpus, the corpus also forming the basis for the Word Frequency Book.

The most familiar meanings of these two words, on the other hand, are probably the following:

carry: to bear in one's hands or arms, on one's shoulders or back, etc., while moving; to transport or convey

carriage: a four-wheeled passenger vehicle, usually drawn by horses

These two meanings are also related, but not as directly as the first two cited. Our semantic code for the relationship between carriage and carry (or between any word and its immediate ancestor) would consist of two digits, the first representing the degree of semantic relatedness between the two most similar meanings, the second representing the degree of relatedness between the two most similar familiar meanings.

Another two-digit code was used to encode the relative familiarity of the meanings represented by the two digits in the semantic code.

There are two further qualifications about the use of the two-digit semantic code. If the two most similar meanings of two words were also familiar meanings, then the second digit was either used to encode the relationship between other familiar meanings of the two words, or else was set equal to the first digit.

For example, the word miserable has as its immediate ancestor misery. It also has two meanings, as in "he made her life miserable" and "miserable weather." Both of these meanings would be considered familiar meanings, the first being perhaps slightly more frequent or salient, and definitely being somewhat more closely related to the meaning of misery. The first digit of the semantic code was used to encode the meaning of

miserable in "he made her life miserable." The second digit was used for the meaning of miserable in "miserable weather."

The analyses reported here, unless specified otherwise, will be based on only the second of the two digits in the semantic code. We feel that the child's experience in learning the meaning of carriage, or figuring out its meaning in context, would be most accurately represented by dealing with the most familiar meanings of the word. It would underestimate the amount of semantic opacity involved in word-formation processes to always measure only the semantic distance between the two most similar meanings of two related words.

Degrees of Semantic Relatedness

The American Heritage School Dictionary was used as the primary reference for determining the meanings of words, since this dictionary is based on the corpus we have analysed, and thus reflects meanings that actually occurred in the corpus. Other dictionaries were also used, primarily to determine the nature and existance of less familiar meanings.

The code for semantic relatedness was defined in terms of the following question: Assuming that the child knew the meaning of the immediate ancestor, but not the meaning of the target word, to what extent would the child be able to determine the meaning of the target word when encountering it in context while reading? The following levels of coding were used:

SEM 0. This indicates that the semantic relationship between target word and immediate ancestor is semantically transparent. There are no semantic features in the target word that are not found in the immediate ancestor, with the possible exception of any semantic features that would to totally predictable from a change in part of speech. For example, if a child knows the word red and has any grasp of the suffix -ness, that child should be able to compute the meaning of the word redness even without any help at all from the context. This is the level of semantic transparency associated with almost all regular inflections. It is also found in many compounds; if one knows the meaning of plankton and burgers, the meaning of the rather novel word planktonburgers is easy to compute, without any help from the context. Many affixes are similarly transparent; knowledge of the word misinterpret should almost guarentee that a person would understand the word misinterpretation.

SEM 1. This code means that the meaning of the target item could be inferred from the meaning of its immediate ancestor with some, but minimal, help from context; almost any context should do. Any semantic components in the target word beyond those in the immediate ancestor, or different from them, would be trivial and predictable even without help from context. For example, the word entertainer may have some connotations of professional or official status beyond the simple meaning "one who entertains," but these are usually associated with the suffix -er, and therefore could be inferred by a reader even without much contextual information.

SEM 2. This code means that the meaning of the target item could be inferred from the meaning of its immediate ancestor with reasonable help from the context; "one exposure learning" would be possible. The target word may contain nontrivial semantic features different from or in addition to the semantic features in the immediate ancestor, but these would require only a general sort of contextual information to be inferred. For example, the word gunner means not just anyone who uses a gun, but normally is used for military personnel with the specific assignment of using or operating guns. Presumably the semantic components specifying "military personnel" would be inferrable from the general context in which the word was used; the context would most likely, for example, rule out an interpretation of gunner as meaning "gunfighter."

SEM 3. This code means that the meaning of the target item included semantic features that were not inferrable from the meaning of the immediate ancestor without substantial help from the context. For example, the meanings of the words copper and head definitely contribute to the meaning of the word copperhead. One could infer that it might mean something like "something with a head made out of copper, or resembling copper, or of the color of copper." Even with a context like "While walking through the woods I almost stepped on a copperhead," however, one could not be sure whether the object in question was a snake, an insect or spider, or perhaps some rare antique copper coin. Even a phrase such as "bitten by a copperhead" wouldn't distinguish between snakes and spiders.

SEM 4. This code means that the meaning of the target word is related to the meaning of its immediate ancestor, but only distantly. The relationship would probably not be apparent without being pointed out, and one would definitely not be likely to guess the exact meaning of the target word if one knew only the meaning of the immediate ancestor. Examples of pairs of words with this degree of semantic relatedness are: vicious/vice, farewell/well, motley/mottle, inertia/inert, or saucer/sauce.

SEM 5. This code is used for a lack of any discernable semantic connection--cases in which the meaning of the immediate ancestor would be of no use in learning or remembering the meaning of the target word. Examples of such relationships are clerical/cleric, groovy/groove, dashboard/dash. (Remember that we are considering only relatively familiar meanings of each of these words.)

Appendix B contains some additional examples of words and their immediate ancestors illustrating each level of semantic relatedness.

In the original coding system, a further distinction was made for levels SEM 1, SEM 2, and SEM 3 between changes in meaning that were metaphorical versus nonmetaphorical changes or extensions in meaning. This distinction was collapsed in the analyses reported here.

Another part of the coding system was used to capture what might be called "semantic specialization"-- that is, cases in which the immediate ancestor might have a range of meanings, and the target word only would relate to one, or a subset of these. (There are also cases in which the

target word might have a range of meanings beyond those found in the immediate ancestor.) Because the semantic relationship between any two words can be very complex, the analyses reported here were limited to the consideration of the relationship between the two most similar familiar meanings, as already mentioned.

Roughly speaking, SEM 0, SEM 1 and SEM 2 can be thought of as semantically transparent relationships; SEM 3 relationships involve significant unpredictable semantic information; SEM 4 is semantically obscure, and SEM 5 semantically opaque.

Types of Words

Estimates of the total number of words in English differ not only in how words are counted--e.g., whether derived forms are counted as separate from their bases or not--but also in terms of whether certain classes of words are counted at all. The WFB contains various special categories of words that are often excluded from counts of words: proper names, numbers, formulae, compounds containing numbers, abbreviations, and nonwords (strings of characters that clearly do not represent vocabulary items). Each item in our sample was marked as to whether it belonged in any of these categories. Details of the criteria used in coding are given in Appendix C.

Unlike some vocabulary researchers, we did not mark words as rare, archaic, obsolete, technical, or scientific names of flora or fauna. If a word actually occurred in the WFB, children do encounter it in their school reading; we consider this a justifiable operational criterion for

defining the boundaries of printed school English. Rather than trying to come up with criteria for specialized or technical vocabulary, we feel that such distinctions, if they become necessary, could be best defined operationally in terms of the actual distribution of words in the corpus.

Results

The result of our coding process was a list of 8,669 items, 7,260 being from the original sample, and the rest added to account for missing ancestors, disambiguations, and second or other members of compounds.[1] Each item on the list has an immediate ancestor, if one exists, and a code representing what type of word it is and the morphological and semantic characteristics of its relationship to its immediate ancestor.

From this list, we can count the number of items falling into each of the word-type and relationship categories in our coding system. Table

---

Insert Table 4 about here.

---

4 gives a summary of the results. For each category, this table gives five different figures. Sample N is the number of items in our sample falling into this category; Sample % is the percent of our sample which this category constitutes, i.e. 100 x Sample N/7,260. The Corpus N is the estimated number of items in this category that would be found in the whole WFB. The Population N is the number of words in the total vocabulary of printed school English (grades 3 through 9) that would fall into this category. Population % is the percentage of words in this category in the population, i.e. 100 x Population N/609,606.

Since our sample is essentially a random sample of the WFB, we can assume that the percentage of items in a category in our sample will be approximately the percentage of items in that category for the entire WFB. However, there is an important sense in which the WFB (and hence our sample of it) is not representative of the population of words from which it is drawn. As the analyses by Carroll, Davies, and Richman (1971) indicate (see Table B-8 on p. xxxvi) all of the roughly 14,000 words in printed school English with frequencies greater than 2.5 per million would be expected to occur at least several times in the WFB. On the other hand, of the more than 200,000 words with a frequency of less than two per billion, less than 100 would be expected to show up in a corpus this size. Thus, in extrapolating from any corpus to the total vocabulary, a very high frequency word represents only itself, so to speak, whereas a low frequency word must be taken as representative of a large number of low frequency words which did not actually appear in the corpus.

Our estimates of the composition of the population have taken this into account by assigning a weight to each word, which is an inverse function of its frequency.[2] This is why the Population % is often substantially different from the Sample %. For example, 11.65% of the words in our sample are morphologically basic. However, it turns out that morphologically basic words are not evenly distributed by frequency. Among the most frequent words in our sample (those that would occur on the average twice or more in a million running words of text) almost 28%

were morphologically basic. However, among the less frequent words this percentage decreased, averaging around 6% in the lower frequency ranges. The percentage of morphologically basic words in the population (7.46%) reflects the fact that the population of words in printed school English has a higher proportion of low frequency words than does the WFB or our sample.

Table 4 is organized as follows: First of all, the different coding categories are arranged approximately according to how they relate to possible definitions of "word." The first group of coding categories are those which would be counted as constituting "separate words" in many definitions of "word," and which would appear as separate entries in most dictionaries. The second group of coding categories are those that might not be considered separate words for some purposes, but would often have separate entries in dictionaries. For example, mice might not always be considered to be a separate word from mouse, for the purpose of counting words, but it would occur as a separate entry in most dictionaries.

The third group of categories contains those such as regular inflections that would not normally occur as separate items in dictionaries.

The fourth group contains categories of proper names, which are excluded from some, but not all, dictionaries and estimates of vocabulary size. Proper names were further subdivided as follows: Basic proper names are those proper names which were also categorized as morphologically basic. Derived proper names are words derived from

proper names by some word-formation process, i.e., by suffixation, prefixation, compounding, or some morphologically idiosyncratic relationship. Inflectional and other varients of proper names include plurals and other varients of proper names that would not be given separate entries in a dictionary. Capitalizations homographic with proper names are those forms, such as Cliff, which might be either a proper name or the capitalization of a non-proper name. Since the noncapitalized form cliff has already been counted elsewhere, we have counted these as constituting proper names. In answer to the question "How many distinct proper names are there?" one would probably want to include all of these categories except for "inflectional and other varients of proper names."

The remaining categories in Table 4 are those which would not normally be counted as separate words or be listed as words in a dictionary.

Note that the categories of special types of words--proper names, formulae and numbers, compounds containing numbers, nonwords and foreign words--are not included in the relationship categories in the first three groups. Thus, the category "morphologically basic words" actually includes only morphologically basic words which are not proper names, foreign words, numbers, etc.

Even without further analysis, certain things are already clear about the estimated vocabulary of printed school English. Most importantly, it is very large. By many definitions of "word," the

population includes over 200,000 words, and another 100,000 proper names. A large number of words--over 170,000--are derived by suffixation, prefixation, and compounding, but there are still quite a few (45,000) which are basic, that is, which cannot be derived from any other word.

The WFB alone contains a vocabulary larger than some estimates of the vocabulary size of average high school seniors--who should presumably be able to read any of the reading material for grades 3 through 9 without too much difficulty.

In Table 5, estimates of the number of derived words in the population are broken down according to relationship type--suffixation, prefixation, compounding, and idiosyncratic relationships--and by degree

---

Insert Table 5 about here.

---

of semantic relatedness. For some purposes we can divide the degrees of semantic relatedness into two classes: SEM 0, SEM 1 and SEM 2 constitue those cases in which the relationship is essentially transparent. A child could, given the meaning of the base, figure out the meaning of the derived form, perhaps with some help from context. SEM 3, SEM 4 and SEM 5, on the other hand, include derived forms whose meanings are not completely predictable from the meanings of their bases, so that they must in effect be learned as separate items.

From Table 5 we see that there are an estimated 139,020 derived forms in the population whose meanings are transparently related to the

meanings of their bases. This suggests strongly that knowledge of word-formation processes opens up vast amounts of vocabulary to the reader. Conversely, a reader who cannot take advantage of morphological relatedness among words has in some sense more than twice as many words to deal with as the reader who utilizes these relationships.

There are also 43,080 derived forms that are relatively opaque semantically. The majority of these, 26,599 words, are at the level SEM 3, which means that although the meaning of the derived form is not completely predictable from the meanings of its component parts, the meanings of the component parts do in fact contribute something to the derived meaning. Even in these cases, then, knowledge of word formation processes will be helpful to the reader trying to figure out the meaning of words in context. On the other hand, however, the semantic opacity of these words is sufficient that many readers--perhaps especially poor readers--will not be able to figure out their meanings, and thus will have to learn them individually.

Table 6 gives the same type of information as Table 5, but computed

---

Insert Table 6 about here.

---

on a slightly different basis. In Table 5, the degree of semantic relationship was based on familiar meanings of derived words and their immediate ancestors. Table 6 is based on the minimal semantic distance between derived words and their immediate ancestors, that is, on the

relationships between the most similar meanings for each pair of words. For example, in Table 5, the relationship between carry and carriage would be counted as relatively opaque, since only the familiar meanings are taken into consideration. For the purposes of Table 6, on the other hand, the semantic relationship between these two words would be counted as transparent, since the most similar meanings were considered. Thus, Table 6 minimizes the number of derived forms that would be considered opaque. Unless otherwise specified, we will use the figures from Table 5 in our discussions of vocabulary composition.

The Number of Webster Main Entry Equivalents

Exactly how many words there are in printed school English depends on the definition of "word" that is adopted. One way to get a meaningful measure is to take as a definition of "word" the criteria for status as a main entry in Webster's Third New International Dictionary, unabridged.[3] This dictionary is of special interest because it was used by Dupuy (1974) as a basis for choosing a set of "basic words" to use in making estimates of absolute vocabulary size. The number of "Webster main entry equivalents" can be computed by including in our count of words the following categories from our coding system (see Table 4 and Appendices A and C): Morphologically basic words, idiosyncratic morphological relationships, suffixation, prefixation, compounding and contractions, truncations, abbreviations, irregular inflections, irregular comparatives and superlatives, alternate forms of words, semantically irregular plurals, "scientific plurals," and derived proper

names. The other categories in Table 4 would be excluded from this count.[4]

Calculated in this way, the numbers of "Webster main entry equivalents" were as follows:

Sample N	3,156
Sample %	43.47
Corpus N	37,707
Population %	39.88
Population N	243,136

How does this compare with the number of words in Webster's Third? Dupuy (1974), on the basis of a very careful count, estimated the number of main entries in Webster's Third to be 240,000. (This number excludes main entries which were prefixes, suffixes, letters and other than first-listed homographs, i.e. it includes only one main entry for each set of homographic words.) However, this estimate is not directly comparable with our estimates of "Webster main entry equivalents," for the following reasons:

1. Our estimates of "Webster main entry equivalents" do not take into account the fact that in Webster's Third, there are separate main entries for regular inflections, comparatives, and superlatives that would fall more than five inches away from their associated main entry in the physical page layout. According to an estimate based on 10 randomly selected pages, about 1.4% of the main entries in Webster's Third, or about 3,360 entries, consist of such regular inflections, comparatives, and superlatives.

2. In Webster's Third, many suffixed forms, mostly in -ly and -ness, are listed as subentries under their associated main entries. According to our estimates, for every 100 entries, there are about 5.02 such subentries. This would amount to 12,048 items in the whole dictionary.

3. Although Webster's Third excludes most proper names, it does include some proper names that would have been coded as basic proper names in our sample. According to Dupuy (1974), there are 23,900 proper names in Webster's Third. On the basis of a small sampling (12 randomly selected pages) we judge that about 31.25% of the proper names in Webster's Third would have been coded as basic proper names in our coding system. This amounts to 7,469 entries.

4. According to Dupuy's estimates, 29.2%, or 70,080 of the main entries in Webster's Third are compound entries; that is, they consist of two or more words separated by spaces, such as heat exhaustion. On the other hand, the corpus of printed material used for the WFB was keypunched in such a way as to exclude such items; with only a very few exceptions, potential compound entries were divided into their component words.

If we exclude from the count of main entries in Webster's Third all entries for regular inflections, comparatives and superlatives, and all basic proper names and compound entries, and if we add to this count the number of suffixed subentries, we have a figure which is directly comparable to the number of "Webster main entry equivalents" in our

estimates for printed school English. The number of main entries in Webster's Third, counted in this way, is 171,139. Thus, somewhat surprisingly, it appears that there are more words in printed school English than in an unabridged dictionary.

One might wonder how this could be. Part of the answer lies in the fact that books in these grade levels sample from a very broad range of topics. Part of the explanation must also lie in the large number of derived words in printed school English. As Table 5 shows, there are about 139,000 semantically transparent derived words, a little more than half of which are compounds. Many of these derived forms, especially the compounds, are low-frequency words coined for specific purposes or contexts, and are not likely to be found in any dictionary. Examples of such words would be essayist-poet, European-owned, ex-florist, and everlengthening. The existance of large numbers of such words in school texts makes knowledge of word-formation processes an important factor in dealing with low-frequency words.

Dupuy's Estimate of the Number of Words in English

Dupuy (1974) undertook not only to construct a vocabulary test, but also to make it a meaningful measure of absolute vocabulary size. Any measure of absolute vocabulary size presupposes a definition of "word;" Dupuy chose to treat vocabulary size in terms of Basic Words, which are defined in terms of the following criteria:

Dupuy took as his source of words Webster's Third New International Dictionary, unabridged. Main entries in this dictionary are "basic words" if they do not fall into any of the following excluded categories:

(1) compound and hyphenated entries,

(2) proper names,

(3) abbreviations,

(4) items which are not main entries in three other dictionaries: The Random House Dictionary of the English Language, The World Book Dictionary, and Funk and Wagnalls New Standard Dictionary of the English Language,

(5) items listed as foreign, archaic, slang or informal, or technical in the Random House Dictionary,

(6) "derived, variant, or redundant" words.

Dupuy estimated that there were 12,300 "basic words" in Webster's Third, by applying these criteria to a representative 1% sample of this dictionary. Using his 123 basic words (the 1% sample of 12,300) as a basis for a vocabulary test, he has estimated vocabulary sizes at different grade levels: 2,000 words in 3rd grade, 4,760 words in 7th grade, and over 7,000 words known by high school seniors.

Initial Comparison of Dupuy's Estimates with Ours. We have already seen, in our estimate of Webster main entry equivalents, that the vocabulary of printed school English is somewhat larger than Webster's Third. (The subset of the vocabulary of printed school English that actually occurs in the WFB is of course smaller, containing a little less than one quarter of the words that are in the unabridged dictionary.) One might expect, then, that the number of basic words in printed school English would be a little larger than Dupuy's estimate, while the number of basic words in the WFB should be substantially smaller.

To compare our estimates of vocabulary size with Dupuy's, we have to determine what would be the closest equivalent in our coding system to Dupuy's Basic Words We will explore this question in more detail below; as an initial basis for comparison, we would compare Dupuy's Basic Words with our category of morphologically basic words. According to our analyses, there are 10,108 morphologically basic words in the WFB, and 45,453 in the population underlying that corpus.

Dupuy (1974) claims to exclude from Basic Words those derived words which are redundant because their "meanings could be understood with knowledge of the meaning of the word and affix." We could therefore add to our count of basic words those derived words with the level of semantic transparency SEM 3, SEM 4 or SEM 5. This would bring the number of basic words in the WFB up to 16,655, and in the population, to 88,533.

On the basis of this initial comparison, Dupuy's figures seem to be underestimates by a substantial degree. His estimate of the number of basic words might be in the ballpark, if it were supposed to reflect the number of basic words a single child of average ability might encounter in school reading material in grades 3 through 9. His sample of basic words was intended, however, to be representative of the entire English vocabulary as represented by Webster's Third New International Dictionary, unabridged. This would lead one to expect that the number of basic words would be somewhat similar to the number we estimated for printed school English.

Sources of the Differences between Dupuy's Estimate and Ours. Having established that Dupuy's estimate of the number of basic words in English is much smaller than would be expected on the basis of our analysis of the words in the Word Frequency Book, we would like to ascertain as closely as possible the reasons for the difference. There are two major possible sources of difference: (a) differences in the corpora used in defining the population of words, and (b) differences in the definition of what constitutes a basic word. It is clear already that factor (a) is not the problem, since the vocabulary of printed school English is slightly larger than Webster's Third. The disagreement between our estimates and Dupuy's must, therefore, lie mostly in the criteria adopted for "Basic Words."

First of all, we want to determine what are the differences between our coding category "morphologically basic words" and his category of Basic Words. To do this, we will look at some of Dupuy's criteria in detail, and, in this process, estimate how many words might be added to Dupuy's estimate if his criteria were adjusted in the directions we will suggest.

Dupuy excludes from his category of basic words certain categories of words that would be included among our "morphologically basic words." Specifically, he excludes items that were not main entries in the four dictionaries he used, and items that were classed as technical, foreign, slang, or archaic in the Random House dictionary.

The first of these categories seems to contain the largest number of words--an estimated 97,900 main entries in Webster's Third are excluded because they did not appear as main entries in the other three dictionaries. A substantial number of these would also have been excluded on the basis of other criteria as well; for example, around half of th items in the list (e.g. abruptly, academician, acknowledgeable) would have been excluded as semantically transparent derivatives.

The motivation for excluding such items is clear, and seems legitimate: A list of the basic words in English should include words that really are English words; and one might assume that any item that is really a word in English would in fact show up in any substantial dictionary. But there are some problems with this principle of exclusion. First, any dictionary (besides the OED, anyway) necessarily excludes large numbers of possible entries, and one cannot assume that the editors' criteria, whatever they may have been, were appropriate for the purpose for which the list of basic words is being compiled.

Second, even a consensus among dictionaries cannot tell us what words actually do occur in the materials children read in school. On the other hand, the American Heritage Intermediate Corpus was carefully selected to be representative of printed materials used in schools in grades three through nine, and gives us a solid basis for an operational definition of what is a word in "printed school English."

Among the words excluded because they were not main entries in all four dictionaries were an estimated 291 that were morphologically basic

(in the sense that they could not be analysed into free or recognizable bound stems). (This estimate is based on an analysis of one-third of the 979 items in this category.) Another estimated 238 items in this group were morphologically, but not semantically, analysable, for example, asthenobiosis, clasmatocyte, hangbird, moosewood. Thus, there could be as many as 500 items among these words that might be counted as basic words under somewhat more liberal criteria. If even a quarter of these were actually counted as basic words, it would double the size of Dupuy's original estimate.

Finally, there are some words among those excluded as technical which seem to be part of general vocabulary: coda, creosol, formaldehyde, herpes, holmium, methyl, orthogonal, and placebo. These 8 words, since they are part of a 1% sample, would add another 800 words to Dupuy's estimate if they were included.

Compound and Hyphenated Entries. Both the WFB and Dupuy exclude all compound entries, that is, items consisting of two or more words separated by spaces. In the case of the WFB, this was due to the methods of keypunching adopted; with only a very few exceptions, words separated by spaces were entered as separate words. (The exceptions were a few compound names such as New York that were incorrectly punched as single items (that is, as NewYork) instead of as separate words.) In the case of Dupuy's analysis, compound entries, although included as main entries in Webster's Third, were excluded from the count of basic words. However, Dupuy also automatically excluded all hyphenated entries, whatever their

nature. Our analysis, on the other hand, treats hyphenated entries as it would compounds (that is, compounds not separated by spaces) or affixed forms. Any such form is individually coded in terms of its semantic transparency. In our estimate of vocabulary size, we would want to include any complex form, hyphenated or not, which would be coded as SEM 3, SEM 4, or SEM 5, that is, which was semantically opaque to the extent that it would have to be learned separately, since its meaning could not be inferred from the meanings of the component parts.

Therefore, in applying our coding system to Dupuy's corpus of words, we want to determine how many of the hyphenated forms excluded by Dupuy are semantically opaque. Of the 775 compound and hyphenated entries excluded from the list of basic words by Dupuy, only 77 are hyphenated. Of these, we would consider at least 22 to be semantically opaque to the extent that they would have to be learned as separate items. These 22 are:

all-fired	cab-over	cap-and-ball
charge-a-plate	chaff-flower	clip-clop
cross-staff	crinkum-crankum	cuckoo-bread
dew-drink	double-talk	dove's-foot
down-and-out	games-all	hokus-pokus
jack-by-the-hedge	last-ditch	man-about-town
poker-faced	rip-rap	small-beer
whing-ding		

To the extent that these do in fact represent items that would have to be learned separately, because their meanings are not inferrable from the meanings of their parts, we would have to add this number of items to Dupuy's estimate of absolute vocabulary size to bring it in line with our criteria. Since Dupuy's estimate is based on a one-percent sample, this means adding 2,200 words to his original estimate of vocabulary size.

Derived, Variant, or Redundant Words. We will continue the comparison of vocabulary size estimates by reviewing the criteria used to exlude from the class of basic words those considered to be "derived, variant, or redundant." In addition to examining the criteria, we will present a reanalysis of the 184 words listed by Dupuy in the "derived, variant, or redundant" category. Dupuy uses the following criteria:

> A main entry was considered a derived or variant word form if in any of the four dictionaries
>
> 1. The definition mentioned or referred back to another form of the same word (e.g., beck: a beckoning gesture) or was simply a different tense form (e.g., supposed: suppose).
>
> 2. The definition was simply a different spelling (e.g., calimanco: calamanco).
>
> 3. The definition was a different word which provided a fuller definition (e.g., boxberry: the checkerberry).
>
> 4. The entry was a combination of two or more words and the definition included a reference to one or more of the words (e.g., bookkeeper: one who keeps account books).
>
> 5. The entry word was a derived form with a base word and affix whose meaning could be understood with knowledge of the meaning of the word and affix (e.g., adiabatic: not diabatic).

For each of these criteria, there are cases in which words will be excluded from the count of basic words which would in fact have to be learned as separate items in the process of vocabulary acquisition.

In the case of criterion 1, there are cases where a different tense form may in fact have meanings divergent enough from its stem so that this meaning would not be easily inferred. For example, striking, imposing, blooming, collected, elevated, and hearing all have meanings which are quite distinct from the meanings of their stems.

In the case of criterion 2, it would in general seem right to count as "the same word" variants that different only in details of spelling. However, there are also cases of variation in spelling, for example draught and draft which are substantial enough to pose real problems to a reader who is familiar with one variant and not the other.

Criterion 3 is probably the most questionable of all, from the perspective of the reader or child learning vocablary. A reader encountering the word milfoil in a text, until he or she turns to the dictionary, is presumably not aided by the fact that this word can be defined simply in terms of another word, yarrow. In fact, if the reader does turn to the dictionary, this type of definition is likely to pose an additional obstacle, if, as is often the case, the word in the definition is as obscure as is the word defined.

Criterion 4 is appropriate if it is applied to words whose meanings can in fact be understood from the meanings of their component parts. In practice, however, Dupuy has used to it exclude from his count of basic words items whose meanings are not all that transparent: fiddlewood, flapdragon, howbeit, leapfrog, seismoscope, silviculture, and threadfin.

Criterion 5, like criterion 4, is appropriate only if the compound item has a meaning that is truly predictable from the meanings of its component parts. Dupuy includes as derived words the following, whose meanings are either not fully predictable on the basis of their component parts, or which rely on relatively rare meanings of their components: chanceful, clamper, coloratura, conquistador, defrock, episcopalism, extravaganza, gymnasiast, provisional, rarefy, and valedictorian.

Applying Our Coding Criteria to Dupuy's Derived, Variant or Redunant Words. Dupuy lists 184 words as derived, variant, or redudant. We applied our coding system to these words to see how many of these words would be considered redundant in terms of our criteria for grouping words.

First of all, five of the words that Dupuy lists as belonging to this category we were not able to find in Webster's Third New International Dictionary, unabridged, the source of all of Dupuy's words: dashen, deconate, padodite, payraceous, and tragedion. We assume that these are due to misprints in the published version of his list; we further assumed that dashen was supposed to be dasheen, and tragedion was a mispelling of tragedian. Otherwise we did not find likely sources in the dictionary for these apparent errors. This leaves us with 181 words to classify.

Of the remaining words, three appeared to be cases of criterion 3, that is, words defined in terms of other words: dasheen (= taro), milfoil (= yarrow), and diesis (= double dagger). As mentioned above, we

would not consider these words to be redundant from the point of view of a reader trying to understand a text, or a child learning vocabulary.

Twelve items from the 181 seem to be alternate spellings (although a few might also be treated as meeting Criterion 3). Listed with their alternate spellings, these are:

bressummer	breastsummer
cullender	colander
draught	draft
ebon	ebony
floatage	flotage
further	farther
hagberry	hackberry
insphere	ensphere
jetton	jeton
koorajong	kurrajong
mediaeval	medieval
proa	prau

Conservatively, draught, and perhaps also proa, are distinct enough in spelling from their alternate forms to present some difficulty to a reader who knew only one form of the word.

The remaining 166 words were coded in terms of the transparency of the semantic relationship between the word and its component parts, according to the same system used in our coding of the sample from the Word Frequency Book.

Defining "semantically opaque" as SEM 3, SEM 4 or SEM 5, there are 43 items among the 184 coded which would be counted as semantically opaque.

In contrasting our criteria with Dupuy's, and applying our criteria to his list of words, we have come up with the following additions to his original set of basic words:

8 words listed in the Random House dictionary as "technical" which we would consider part of general vocabulary.

291 (estimated) morphologically basic words among those Dupuy excluded because they did not occur as main entries in all four of the dictionaries he used.

238 (estimated) words among those excluded because they did not occur in all four dictionaries, which were morphologically complex, but semantically opaque.

22 semantically opaque hyphenated entries.

3 items counted as "redundant" by Dupuy (dasheen, milfoil, and diesis) which we feel would have to be learned as separate items.

2 difficult spellings (draught and proa) so different from their alternative forms that they would presumably require separate learning.

43 words counted as redundant by Dupuy, which we consider to be semantically opaque.

This adds up to a total of 607 additional words beyond the 123 already counted as basic by Dupuy. This would bring the total number of basic words in Webster's Third up to 73,000. This figure is much closer to our estimate of basic words in printed school English (88,533); although it is still a little lower than our figure, it is almost six times as great as Dupuy's original estimate of the number of basic words.

The bulk of the difference between Dupuy's original estimate and our figures seem to be traceable to two main factors: First, Dupuy's use of four dictionaries excludes a large number of words--most of them rather low in frequency to be sure--which we would include. Second, he clearly sets a different cut-off point with respect to which words are to be counted as semantically redundant. He seems to place a much greater weight on morphological relatedness, and considers as redundant words which we would consider to have only rather distant semantic relationships.

In summary, we might say that Dupuy has adopted a prescriptive rather than descriptive concept of what constitutes a basic word in English, and that his estimates do not at all reflect the diversity of vocabulary encountered by children in reading school texts.

Seashore and Eckerson's Estimate

Like Dupuy (1974), Seashore and Eckerson (1940) attempted to construct a test which would measure not only relative vocabulary knowledge, but also given an indication of the absolute size of a person's vocabulary. They also used the method of selecting a random sample of items in an unabridged dictionary. We want to contrast our estimates of vocabulary size with theirs first, because their study has served as a basis for much subsequent research in vocabulary size, and second, because it has been subject to careful scrutiny by Lorge and Chall (1963).

Seashore and Eckerson took as their population of words the entries in Funk and Wagnalls' New Standard Dictionary of the English Language, the two volume edition of 1937. This dictionary was chosen because it was large enough to represent the full range of adult vocabulary without including extremely rare words. Also, it contains all words in a single alphabetical order, making it easier to construct a subsample for testing.

This dictionary contains two types of entries: "basic" words, or main entries, printed in heavier type and next to the left margin, and "derivative" terms, which are indented under the basic term. Seashore and Eckerson estimated that the dictionary contains 166,247 "basic" words, and an additional 204,018 "derivative" words, excluding multiple meanings and variants in spelling.

To some extent, the distinction between basic and derived entries can be stated in terms of word formation processes. That is, derivative entries are words derived from their basic entries by suffixation or compounding. Seashore and Eckerson give the example of the basic word loyal and its derivatives Loyal Legion, loyalism, loyalize, and loyally. However, not all words derived by compounding or suffixation are listed as derivatives; many such items are basic words. For example, master, masterful, masterhood, masterless, masterly, masterpiece, mastership, mastersinger, masterwork, and mastery are all basic words, that is, main entries in Funk and Wagnalls' dictionary. Furthermore, prefixed forms, because they occur elsewhere in an alphabetic list, also constitute separate main entries.

The criteria for placement of an item as a main or derivative entry are not explicitly given in the dictionary. The principles followed seem to be approximately these: First, compound entries (that is, entries with internal spaces) are treated as derived entries, except in the case of a few which are also proper names. Second, suffixed items whose meaning is predictable from that of the basic word with no or little additional definition are usually treated as derived entries. This includes most adverbs in -ly, nominalizations with -ness, and many other adjectival forms. For the remaining suffixed items and compounds, which could be listed either as basic or derivative, one of the criteria for placement seems to be some notion of "importance." For example, iceboat and icebreaker are basic entries, while icecliff, icefoot, icequake, and others are listed as derivatives. "Importance" seems to correspond pretty closely to frequency.

In some cases, alphabetical order and the arrangement of words seem to play a role. For example, under the basic item Eurystomata are listed the derived words eurystomatous, eurystoman, eurystomous, eurystome, eurythermal, and eurythermic. Were t to precede s in the alphabet, it seems likely that eurythermal would have been the basic word, and Eurystomata one of the derivative items. The principle followed here seems to be that if a number of relatively rare or unimportant compounds occur in succession, the first is given as a main entry, and the following as derivatives. This also seems to be the case, for example, when under the basic word meteoromancy are listed the derivative items meteorometer, meteoroscope, and meteoroscopy.

A slight further complication is that some compounds are listed as derived items, and also as main entries, with the main entry referring to the definition given for the derived item.

In many cases, derived items are redundant, or semantically transparent. That is, if one knows, for example, the meaning of the basic item evangelical, the meaning of the derivative evangelicalism is likely to be self-evident. On the other hand, a substantial proportion of the derivitive entries in Funk and Wagnalls may not be so semantically transparent. For example, knowing the meaning of stay does not guarentee that one will be able to figure out the meaning of stayplow (a type of plant, also called restharrow).

It cannot be assumed that all basic entries are semantically distinct, either. For example, one might consider the meaning of gusty as rather obvious, given the meaning of the word gust. Similarly, evaporate, evaporation, and evaporator are listed as distinct basic entries, despite their clear semantic relatedness.

Thus, it is not clear exactly how Seashore and Eckerson's estimates of vocabulary size should be interpreted. The figure of 166,247 basic words and 204,018 derived words, totalling 370,265 words, reflects the make-up of an unabridged dictionary, but cannot be directly interpreted in terms of any particular theory of words and how they are learned.

Lorge and Chall's Critique of Seashore and Eckerson. Lorge and Chall (1963) have critically examined the work of Seashore and Eckerson, and noted several weaknesses. One relates to the problem of space

sampling. The method used to obtain a sample of words from the dictionary--taking the third basic word in the first column of every left-hand page in the dictionary--turns out to produce a sample that is biased in that it contains disproportionately many common or easy words. This makes the vocabulary test based on this sample easier, and hence leads to an overestimation of the vocabulary size of the person taking the test.

Lorge and Chall also noted some errors or inconsistancies in counting. For example, Seashore and Eckerson claimed not to count duplicate spellings in their count of basic words, but Lorge and Chall found that 2% of the basic words in their initial estimte of vocabulary size were in fact duplicate spellings. Another inconsistancy relates to homographs. Lorge and Chall argue that since Seashore and Eckerson take as a criterion of word knowledge recognition of any common meaning of a word, they should not count homographs as separate items. However, homographs (counted as distinct items) amounted to 9% of the basic words in Seashore and Eckerson's estimates.

More importantly, Lorge and Chall disagree with Seashore and Eckerson as to what should be counted in an estimate of vocabulary size. They suggest excluding the following categories of items, which amount to an estimated 30% of the entries in Funk and Wagnalls: Names of persons, Biblical names, other names (mythical, races, etc), names of flora and fauna, geographical place names, abbreviations, suffixes, prefixes, and combining forms. Taking all these adjustments into account, Seashore and

Eckerson's estimate of 166,000 basic words is reduced by about 40%, to 99,600.

Comparison with Our Estimate. How many words are in printed school English if one adopts the criteria from Seashore and Eckerson (1940)? To compute the number of "basic words" by their definition, we can start with our number of "Webster main entry equivalents," and make the following adjustments: First, all but the most common compounds would be excluded, since they would be derived entries in Funk and Wagnalls. Also excluded would be all semantically transparent suffixed forms. On the other hand, we would have to add to our estimate basic proper names and capitalizations homographic with proper names, since these would be main entries in Funk and Wagnalls. (To come up with an estimate based on Lorge and Chall's (1963) revision of the criteria for "basic words," we would exclude these last two categories.) The number corresponding to Seashore and Eckerson's "total words" would be the number of "Webster main entry equivalents," including all derived and compound forms, plus basic proper names and captialization homographic with proper names.

Table 7 compares Seashore and Eckerson's (1940) estimates of the number of words in English with the results of applying comparable

---

Insert Table 7 about here.

---

definitions of "word" to the WFB and the underlying population of words. This table also includes estimates of the number of main entries and

"basic words" in Webster's Third by Dupuy (1974) and the results of applying somewhat similar definitions of "word" to the data in the WFB.

It is interesting to note that in every case but that of Dupuy's "basic words," the authors' original estimates are rather close to the figures derived by applying comparable criteria to the population of words in printed school English. This is an indication that the three sources of vocabulary--printed school English as sampled in the WFB, Webster's Third (unabridged), and the Funk and Wagnalls dictionary used by Seashore and Eckerson (1940)--are all of approximately the same size, especially when adjustments are made for the fact that Webster's Third, unlike the other two sources, includes only a restricted range of proper names, and for the fact that the WFB, unlike the two dictionaries, does not have separate entries for compound items. The differences between the columns in Table 7 are therefore due largely to differences in the definitions of "word" or "basic word" that were adopted. Had the authors been able to agree on these definitions, there would have been fairly close agreement as to the total number of words in English.

How Many Words Are There In English?

In the estimates of total number of words in English we have just been comparing--based on large unabridged dictionaries and a statistical projection to the total vocabulary of printed school English--the major difference between the magnitudes has been due to disagreements about criteria used for counting. To answer the question "How many words are there in English?" one has to determine what is the appropriate definition of "word" to use.

We feel that the best way to approach the counting of words is in terms of distinct word families, where a "word family" is a group of morphologically related words such that if a person knows one member of the family, he or she will probably be able to figure out the meaning of any other member upon encountering it in text, with information from context that would be available for most occurrences of that word.

Counting as distinct word families all morphologically basic words and semantically opaque (SEM 3, SEM 4 and SEM 5) derived words, we have estimated that there are 88,533 distinct word families in printed school English. However, some substantial qualifications must be made before this number can be correctly interpreted.

First of all, how words are to be counted depends on why you are counting them. Our interest in estimating the number of words in printed school English is to determine the size and nature of the task that children face in learning the vocabulary of school texts. Whether we should count understand and misunderstand as one word or two depends on how children actually deal with them. If children who know the meaning of understand can learn the word misunderstand, or interpret it in context, with little or no additional effort, then we would want to count these two words as being members of a single word family.

Therefore, any criterion for counting words must be relative to some level of morphological knowledge. For this reason, a truly meaningful estimate of the number of words in printed school English will require empirical studies of children's knowledge of morphology. Our system of

coding different degrees of semantic relatedness is an attempt to approximate what we believe the results of such studies would be; but it remains speculative until these coding categories can be tied to particular age and ability levels.

Our estimate of 88,533 distinct word families assumes that children in grades 3 through 9 would not be helped much by morphological relatedness among words if the degree of semantic relatedness were SEM 3, SEM 4 or SEM 5. For example, knowing the meanings of hook and worm would not provide sufficient information for the child to guess the full meaning of hookworm unless the context were rich enough to give unmistakable clues for the remaining semantic components (e.g. parasitic, causing disease). Therefore, hookworm and similar derived forms were counted as constituting separate word families. However, if we could somehow establish that 9th graders were able to make use of SEM 3 relationships in learning or interpreting new word meanings, our estimate of the number of distinct word families for ninth graders would have to be reduced to 61,934. Conversely, if we were to find that children at a certain grade level were less adept than we expected at seeing and utilizing relationships among words, our estimate of the number of distinct word families for children at that grade level would have to be revised upwards.

Other Categories of Nonredundant Words. Another way to talk about word families is in terms of redundant versus nonredundant words. If a child who knows the meaning of estimate can automatically interpret or

learn overestimate, the latter word is redundant; it does not contribute to the child's vocabulary learning task, or add to the vocabulary load of a text the child is reading. Our figure for the total number of distinct word families is supposed to reflect the number of nonredundant words in printed school English. However, there may be several types of words not included in this count which also should probably be counted as nonredundant in terms of the effort they would require to learn or interpret.

For example, abbreviations were not included in our count of distinct word families, because they do not constitute distinct words in the prototypical sense. One might consider them to be redundant in that an abbreviation has the same meaning as the word for which it stands. However, the relationship of an abbreviation to its unabbreviated form, and hence its meaning, is not at all obvious in all cases; most often, an abbreviation must be learned as a separate item.

On similar grounds, one might want to include in the count of distinct word families other categories in our coding system such as truncations, irregular inflections, irregular comparitives and superlatives, some alternate forms of words, and semantically irregular plurals. For each category, it could be argued that many or most of the items were not redundant--that is, that knowledge of other, related forms would not guarantee the reader a fair chance of understanding that item when encountering it the first time in reading.

All the categories just mentioned would add only an estimated 4,935 words to the population, bringing our total vocabulary estimate up to 93,468 distinct word families. However, if we want to estimate the total number of words in printed school English in terms of nonredundant items to be learned several other categories of items might be added which would increase this overall figure substantially.

Proper Names. Both Dupuy (1974) and Lorge and Chall (1963) exclude proper names from their count of basic words. This exclusion is presumably based on the fact that proper names are functionally distinct from other vocabulary items in a number of ways. In some theories of meaning, for example, it is argued that proper names have reference, but no meaning, unlike common nouns which can have both reference and meaning. In the context of reading, it might be argued that a child only has to recognize a proper name as being such, and that any information about the individual associated with that name will either be supplied in the story itself, or should be considered knowledge about the world, and not vocabulary knowledge as such.

This is a complex issue, more so than we could do justice in the scope of this paper. One could argue, however, that there is at least a subset of proper names that should be counted as part of general vocabulary. Certainly, the names of characters are usually assigned a referent within the context of a story, so that the reader often needs little, if any, prior knowledge about that name to successfully comprehend the text. But there are some proper names which are most

often not explained within texts, and which the reader must be familiar with in order to properly understand the text. This is certainly true of many familiar geographical place names. Lack of knowledge of the reference of words such as Washington, Florida, Alaska, or Panama could contribute to comprehension failure in exactly the same way that ignorance of the meaning of other words in the text might. Thus there is at least a subset of proper names which on practical grounds might be considered as an integral part of a person's vocabulary knowledge.

A related point is that the line between proper names and other areas of vocabulary--for example, names of flora and fauna, or technical terms-- is not clearly defined. For example, eagle is counted by Dupuy as a basic word, but Megaloceros as a proper name. There are differences between these two words, in terms of usage and frequency, but it isn't clear that these differences bear directly on the classification of an item as a common or proper noun.

Determining which or how many proper names should be included in an estimate of vocabulary size would require some more detailed work on the role of proper names in reading comprehension. A rough estimate, however, was made in the following fashion: Of the 929 morphologically basic proper names in our sample, a count was made of those which intuitively seemed to be "important"--that is, knowledge of them would be likely to be assumed in at least a large proportion of school texts. Eighty proper names met this criterion. A second count, of those proper names that were listed in the American Heritage School Dictionary, gave

the same result. It would seem reasonable to assume that those proper names which were necessary for understanding school texts would be defined in this dictionary, and vice versa.

Since there are eighty proper names in our sample knowledge of the meanings of which would probably be assumed in most school texts, there would be about 956 such names in the WFB. Assuming that important proper names are relatively high frequency words, there would be perhaps 1,000 such names in the population, and possibly several times as many. Especially in the higher grades, one would expect that an increasing number of proper names would be assumed rather than explained in school texts, and thus should be counted as part of the demands on the child's vocabulary knowledge.

Homographs. Most estimates of vocabulary size, and all of those we have been discussing, lump together all homographs. But a child who knows only the noun bear (= animal), when confronted with the verb bear (= carry) in a text for the first time, is encountering a brand new word. Knowledge of the one meaning of bear is no help in figuring out the new meaning. In fact it is probably a hinderance. For this reason, if an estimate of vocabulary size attempts to reflect the number of nonredundant items a child would have to learn, it would have to count distinct meanings of homonyms as separate items. Even related, but somewhat different, meanings of the "same word" may present difficulties to young readers.

An estimate of the extent of homophony in printed school English was made by counting the number of distinct meanings for a random sample of 156 of the morphologically basic words identified in our 7,260-word sample of the words in the Word Frequency Book. The primary dictionary used for determining number of meanings was the American Heritage School Dictionary. Since this dictionary was based on the American Heritage Intermediate Corpus, which also formed the basis for the WFB, it should reflect the number of meanings actually occurring for a given item in that corpus. For words which did not appear in this dictionary, we used Webster's Third New International, unabridged. This introduces a potentially confounding factor, since an unabridged dictionary would be likely to include a larger number of meanings for any given item. However, for each item, a code was used to represent which dictionary was used to determine the number of meanings, so that this could be taken into account in statistical analyses. Morphologically basic words appearing in neither of these two dictionaries were assumed to have only one meaning.

The number of distinct meanings for each word were counted at each of five levels of semantic distinctness, defined in terms of the levels of semantic distance between meanings used in our coding system. One example should make the relationship between the two codes clear: Two meanings are counted as distinct at level SEM 2 if the distance between them was greater than SEM 2 in terms of our original coding system. Two meanings were collapsed (counted as nondistinct) if they were related at a level SEM 2 or lower.

The end points of our scale are defined as follows: At level SEM 0, any variations in meaning listed in the dictionary, however minor, were counted as distinct, along with any meanings for subentries such as other parts of speech, idioms, and phrases. At level SEM 4 two meanings were counted as distinct only if there was no relationship at all between them that would be of any use in learning or remembering the two meanings.

In addition to these five levels, for each word we also encoded the number of homographs, as numbered with superscripts in the American Heritage School Dictionary, or the number of etymologically distinct sources in Webster's Third. A seventh number represented the sum of all phrasal or idiomatic entries associated with each word.

As an example of how this coding system worked, here is how the word desert was analysed. The entries for desert in the American Heritage School dictionary were as follows:

desert(1) n. A dry, barren region, often covered with sand, and having little or no vegetation

adj. Uninhabited: a desert island

desert(2) v. 1. To forsake or leave; abandon

2. To leave (the army or an army post) illegally and with no intention of returning

desert(3) n. Often deserts. That which is deserved or merited

There is a total of five distinct meanings listed in these definitions; thus, the number of distinct meanings at level SEM 0 would be five. At level SEM 1, the two meanings of the verb (desert(2)) would be grouped together, since most contexts should make the military implications of the word desert fairly obvious. At level SEM 2, these four remaining meanings would still be distinct, but at level SEM 3, where any clearly related meanings are grouped together, the adjective meaning of desert(1) would be grouped with the meanings of desert(2). At the level SEM 4, the meaning of desert(1) (the noun) would be grouped together with these, leaving only 2 distinct meanings. This word would still be counted as three homographs, based on the numbering system of the American Heritage School Dictionary.

One might argue that the noun meaning of desert(1) should have been grouped with the verb meanings at level SEM 3 instead of SEM 4, since the relationship between the two is fairly clear. On the other hand, perhaps due to the difference in pronunciation, we would guess that most individuals do not make a conscious connection between the two meanings.

Ultimately, such decisions would have to be based on empirical studies. On the other hand, while our current coding system is subjective, Dupuy's (1974) criteria for whether or not a word is redundant are not inherently any more objective than ours. Our criteria have the advantages of making finer distinctions, that is, recognizing degrees of semantic transparency, and being at least in principle defined in terms of the difficulty a word might present to children encountering

it for the first time in reading. In addition, the two end points of our scale of the number of meanings for a word (SEM 0 and the number of homographs) are operationally defined.

The results of this analysis are presented in Table 8. For each measure of polysemy--the five levels of semantic distinctness, the number of homographs, and the number of phrasal and idiomatic entries, two measures are given.

Insert Table 8 about here.

The first is the mean number of meanings; that is, the total number of distinct meanings divided by the number of morphologically basic words. We can assume that our sample of 156 morphologically basic words is representative of the morphologically basic words in the WFB. The frequency distribution of morphologically basic words in the population is different than that in the WFB. For levels SEM 2 and SEM 3, estimates are given for the population as well, taking into account that the population will have a higher proportion of words with lower frequencies and fewer meanings. (Estimates are given for levels SEM 2 and SEM 3 because these levels are most likely to reflect the knowledge of relatedness among word meanings in grades 3 through 9. In our opinion, SEM 3 should give a very conservative estimate, and probably an underestimate, of the number of meanings that would be functionally distinct at this level.)

The second figure is the total number of distinct meanings among the morphologically basic words. Estimates are given for the WFB, and, for levels SEM 2 and SEM 3, for the underlying population as well. There are an estimated 10,108 morphologically basic words in the WFB. At level SEM 2, there are about 2.038 distinct meanings per morphologically basic word, and hence a total of 20,600 distinct meanings of morphologically basic words. For the population of morphologically basic words in printed school English, there would be approximately 73,417 distinct meanings. These figures are lower for level SEM 3, since fewer meanings are counted as distinct at this level.

A count of all semantically distinct vocabulary items will have to include not only all meanings of morphologically basic words, but also meanings of semantically opaque derived words. (Numbers for these are taken from Table 6, which gives a more conservative estimate of the number of semantically opaque forms, assuming, so to speak, that the individual already knows all the meanings of the base forms.) This measure can be added to the number of distinct meanings among the morphologically basic words to give an estimate of the total number of distinct meanings in the vocabulary (for any given criterion for semantic distinctness).

Table 9 gives the total number of distinct meanings at two levels of semantic relatedness. At level SEM 2, the total number of distinct meanings in printed school English is estimated at 105,238. At level SEM 3, the total is 67,417.

Insert Table 9 about here.

Compound Entries. Dupuy (1974) and the Word Frequency Book both exclude compound entries, that is, those which consist of two or more words separated by spaces. Approaching the issue of vocabulary size from the perspective of learning new items, it would seem more appropriate to exclude those (and only those) compound entries whose meanings were computable on the basis of the meanings of their parts, so that a child encountering this combination for the first time in the process of reading could, with a little help from context, infer its meaning.

A survey of the 698 compound entries excluded by Dupuy indicates that a substantial number of them have meanings which are not totally predictable from the meanings of their parts. First of all, there are idioms such as bum steer, favorite son, one-night stand, or straw man. There are about 77 such items among the 698 excluded by Dupuy which have meanings obscure enough that a child would almost undoubtedly have to learn them as separate items.

There are at least 134 additional items which are semantically opaque in the following sense: It is clear that a snake fly is a kind of fly, or that a snap bean is a kind of bean. But the word snake does not really tell what kind of fly a snake fly is; nor does the word snap give enough information, on the basis of its literal meaning, to distinguish snap beans from other beans. The actual reference of such terms must be

learned individually for each such item. Altogether, then, there are 211 items among the 698 compound terms excluded by Dupuy which are idiomatic in that their exact meaning is not predictable from the meanings of their component parts.

Since Dupuy's analysis is based on a 1% sample of Webster's Third, this means that there are approximately 21,100 semantically opaque compound items in that dictionary. Considering that the vocabulary of printed school English has been found to be comparable to that in an unabridged dictionary in other respects, we would expect somewhere near this number of semantically opaque compound items to be found in school texts as well. Much of this number, however, has already been incorporated into our measures of polysemy, since our count of the number of distinct meanings included all phrasal and idiomatic entries related to any morphologically basic word. From the number of semantically opaque compound entries in Webster's Third, however, we can be fairly sure that our estimate of the contribution of polysemy to the size of vocabulary is a conservative one.

Total Count of Nonredundant Items.

Given an estimate of at least 1,000 proper names that should be counted as part of general vocabulary knowledge, and 4,000 abbreviations, irregular inflections, and other orthographically nonredundant words, added to the figures already calculated for incorporating polysemy, we come up with an estimate of 110,000 distinct words in printed school English. This number assumes that individuals are only able to utilize

SEM 0, SEM 1 and SEM 2 relationships in learning or interpreting new words. For someone who is able to utilize SEM 3 relationships as well, the number of distinct words would be 72,000.

The Distribution of Words by Frequency

So far, we have shown that printed school English includes a very large number of words, comparable to the number of words in a fairly large unabridged dictionary. Now we would like to determine, as far as is possible, how many of these words a student in grades three through nine might actually encounter in reading, and how many of these words would actually be useful to a student.

One way to approach this question is to look at the frequencies of the words. Table 10 shows how the words in printed school English are distributed by frequency. Frequencies are given in terms of U, or estimated frequency per million words of text. A word with U = 10.0, for example, would be expected to occur on the average about ten times in a million words of text. Details of how U is calculated are found in the WFB (p. xl).

The numbers of graphically distinct types with a frequency equal to or greater than a given value are interpolated from tables in the WFB. These numbers are predicted on the basis of the lognormal model; according to this model, if frequencies are expressed logarithmically, words will be found to occur in a normal distribution along the frequency continuum.

Insert Table 10 about here.

The number of morphologically basic words and semantically opaque derivatives (included here are SEM 3, SEM 4 and SEM 5 derived forms) gives us an approximate idea of the number of distinct word families among the words above any given frequency level. It should be cautioned that the number of distinct word families at any given level is underestimated somewhat, since the most frequent member of a word family is sometimes a regular inflection or transparent derived form. The word month, for example, has a U of 71.635, whereas the U of the plural months is 115.15. Thus, the word family containing month and months is not included in the count of 555 morphologically basic words and semantically opaque derivatives that have a U of 100.0 or greater. However, among the words in that frequency range, one does encounter a representative of the month family, so that more than 555 word families are actually represented.

Semantically transparent derivatives include those derived words (suffixed, prefixed and compound forms, and a few idiosyncratic forms like prophesy), the meanings of which are largely or wholy predictable from the meanings of their component parts (i.e., SEM 0, SEM 1 and SEM 2).

At least two things are clear about the distribution of words by frequency. First of all, most words are in the lower ranges of the

frequency spectrum. About half the words in printed school English, no matter how one counts them, occur roughly once in a billion words of text or less. Second, semantically transparent derivatives are skewed towards the low end of the frequency distribution to a greater degree than are morphologiclly basic words and semantically opaque derivatives. The relative proportion of these two categories changes radically from one end of the distribution to the other; although there are substantially more transparent derivatives than there are morphologically basic words and semantically opaque derivatives, among the most frequent words the semantically transparent derivatives are relatively rare.

This difference in distributions has some distinct implications for instruction. If a child were exposed only to vocabulary controlled carefully by frequency, there would be both relatively little opportunity to learn, and little necessity to make use of, the word-formation processes that relate derived words to their component parts. The relatively few transparent derived words that do occur in the higher frequency ranges are likely to be learned, at least at first, as unanalyzed wholes (cf. Kuczaj, 1977; Silvestri & Silvestri, 1977). On the other hand, it is clear that as one's exposure to the language expands into the lower frequency ranges, knowledge of word-formation processes becomes an increasingly necessary skill.

At this point it might be appropriate to comment on the importance of low frequency words. One might be tempted to argue, after all, that words occurring one in a million words of text or less--however many such

words there may be--are really not worth much consideration. If the student encounters such words on the average once a year or less (for any individual word) there wouldn't seem to be a need to include them in any program of vocabulary instruction.

But before jumping to any conclusions about words in the lower ranges of the frequency continuum, it might be useful to look at what words are actually involved. Many of them do seem to be of little general use, but there are some rather useful-seeming words there as well. Among the words occurring less than once in 100 million words of text (U = 0.008) are ones such as:

amnesty	elevate	gnome	persecute
appall	evict	hornswoggle	racoon
assimilate	expound	ignoramus	rambunctious
busybody	flex	jellybean	rote
cheeseburger	fluent	liturgy	shamrock
contemporary	fume	mediate	stenographer
eczema	furor	papaya	syncopate

Among the even rarer words, occurring less than three times in a billion words of text (U = 0.0025) are:

ammeter	anneal	billfold	cloverleaf
cyanide	deform	hex	orthographic
solenoid	template	unwieldy	ventilate
calliope	emanate	extinguish	flippant
nettle	pidgin	saturate	seagull
spinnaker	fresco	inflate	sacrament

This is not a representative sample of low-frequency words, to be sure, but these examples do demonstrate that just because a word has a relatively low frequency in printed school English does not mean that it is of little utility.

Since a word's frequency does correlate with the probability that an individual will know that word, it is easy to mistakenly identify low frequency with difficulty. But almost any book by Dr. Seuss will serve as proof that utterly novel words are not necessarily difficult for a child to read. Yet many such words occur only once in a single story, and thus would have astronomically low frequencies in any large scale survey of word freqency.

The frequency of a word reflects a number of factors; one of them is often the conceptual difficulty of the word. But in general it might be said that a word's frequency reflects the range of contexts in which the word might appear. A "rare" word such as sacrament is important within a certain set of contexts, but this set of contexts is very small compared to the universe of contexts that are covered in printed school English.

It should also be noted that frequency studies such as the WFB that involve very large samples of written language are not representative of an individual student's exposure to the language. Because choice of words will be more consistent within a given author's works or a given subject category, any individual student will not get a random sample of vocabulary containing a wide range of low frequency words occurring once each. Rather, in a given student's reading, most low frequency words will not occur at all, and of those that do, many may occur a number of times.

There is an important sense in which the frequencies listed in the WFB underestimate the true frequency of occurrence for a given word. A

student's exposure to the word drive, for example, is not a function of the frequency of that graphically distinct type alone, but rather, a function of the sum of the frequencies of all members of the family. In this case, one would certainly want to include forms such as Drive, driven, driver, Driver, driver's, Driver's, drivers, drivers', drives, and drove. The frequency of this entire family is over three times greater than the frequency of the morphologically basic word drive. This particular family is more extensive than many, but it is still true that family frequency is always greater than or equal to the frequency of any individual member. In this sense, students may encounter some of the low-frequency words in printed school English more often than one would gather from the frequencies reported in the WFB.

Finally, it should be noted that the materials on which the WFB is based tend to have a higher proportion of high frequency words than does printed matter written for adults. This means that the frequencies reported for rare words in the WFB will in general be lower than the reported frequencies for the same words in adult materials.

The distribution of words by frequency does show that of the many words in the vocabulary of printed school English, a large portion have very low frequencies. Nevertheless, one must be careful in interpreting this fact. It would be a mistake to suppose, for example, that all words occurring once in a million words of text were so technical or specialized as to be of no pedagogical significance.

How Many Different Words Do Children Actually Encounter?

To get an accurate picture of the vocabulary that students actually encounter in printed school materials will require both information on the amount and type of reading done by children in and out of school, and a reanalysis of our data by grade level. Our plans for future research include both these steps; at present, however, we can get at least an approximate idea of the number of words students have to deal with in school reading. At the lower end of the spectrum, one might imagine a less able reader at one of the lower grade levels reading as few as ten pages a day from books with large print and frequent pictures, averaging 100 words per page. If this rate were maintained through 100 days of the school year, 100,000 running words of text would be covered. This figure would seem to be a lower limit to the amount of reading done in school between grades three and nine. On the other hand, it does not seem unlikely that an average reader in seventh grade might spend fifty minutes a school day in actual reading, at a rate of 100 to 200 words per minute. In 100 school days, 500,000 to 1,000,000 running words of text would be covered. This is certainly not a maximum; given a higher reading speed, a little more time spent in reading, and more consistent reading during the year, and a child might cover 10,000,000 running words.

The forgoing estimates may be conservative. Carroll (1964) has conjectured that college students may be exposed to as many as a million running words a week in their reading, lectures, and conversations. Our

own conjecture is that there are avid readers from the middle grades who approach this figure.

From the WFB (see Table B-9, p. xxxvii) it appears that a student in grades three through nine who reads 500,000 to 1,000,000 running words of text in a year will be exposed to between 20,000 and 40,000 graphically distinct types. From our analyses of the WFB, this would mean that somewhere between 4,000 and 10,000 distinct word families might be encountered. More precise estimates will require analysis of our data by individual grade levels. In the meantime, we can be fairly confident that an average reader in the upper half of the grade range would encounter at least 5,000 distinct word families in a year, perhaps as many as 10,000. At least 1,000 of these would be families that had not been encountered in the previous year, and it is quite possible that an active reader in these grades could come across three or four thousand totally new vocabulary items in the course of a school year.

Further analyses will allow us to specify with much more precision the number of new word families that a child in any grade would be likely to encounter. However, even the present rough estimates are sufficient to demonstrate that direct instruction could not cover more than a small fraction of the words that a student will actually encounter in school reading.

Word Families in School English

How much interrelatedness is there among words in printed school English? One way to approach this question is in terms of the size of the average word family. If there are are 609,606 graphically distinct types in printed school English, and only 88,533 distinct word families, one would expect there to be 6.88 members per family. This figure is inaccurate, however, because there are several kinds of words (e.g., numbers and proper names) which were not included in any family at all.

Table 11 represents the average composition of a word family in printed school English. Since the concept "word family" can be defined only with respect to some level of morphological ability, we have decided to give figures based on two different definitions.

Insert Table 11 about here.

Definition A adopts a conservative estimate of the number of distinct word families in printed school English. Assuming, in this case, that some individuals might make effective use of even SEM 3 and SEM 4 relatedness in learning derived words, we count as distinct word families only morphologically basic words and derived words with a semantic relatedness level of SEM 5. By this definition there are about 54,000 distinct word families. Since people frequently learn words without perceiving relationships that do exist betwen them (e.g., basement and base) we would consider this to be an underestimate of the

true number of distinct word families; however, it can serve as a reasonable lower limit.

Definition B is the definition of word family we have adopted up to now; it includes morphologically basic words and derivatives at levels SEM 3, SEM 4 and SEM 5. By this definition there are around 88,500 distinct word families. This is by no means an upper limit; as discussed above, the number could be raised considerably if, for example, distinct meanings were counted as separate word families, or if even a small portion of proper names were included. But given that we want a figure comparible to Definition A in excluding proper names and not considering problems of polysemy, this can be taken as our best estimate of the number of distinct word families, for children who can make use of English derivational morphology when the semantic gap between derived word and base is relatively small.

Table 11 shows that for each word known most people will readily interpret .87 to 1.42 words that differ only in minor details of form, and from 1.16 to 1.90 words which are inflections of the base word. It can also be seen that in the average word family, for each base word, there are between 1.57 and 2.57 additional semantically transparent derivatives. For the child who is able to make use of SEM 3 and SEM 4 derivatives, for each word learned there are more than three derived words with meanings recognizably related to that of the base, and at least two of these involving fairly transparent relationships. This demonstrates that the ability to utilize morphological relatedness among

words puts a student at a distinct advantage in dealing with unfamiliar words.

Summary and Implictions

Measures of Absolute Vocabulary Size

Our basic finding is that when a psycholinguistically and pedogogically justifiable way of counting words is employed, the number of words in printed school English is extremely large. Furthermore, our findings imply that previous low estimates of individual vocabulary sizes are in error. Specifically, Dupuy (1974) substantially underestimated vocabulary size because he underestimated the number of basic words in English.

Dupuy (1974) calculated the number of basic words in English for the purpose of creating a vocabulary test that would indicate an individual's total vocabulary size. This test, the Basic Word Vocabulary Test, is advertised as "the only test on the market that yields an estimate of a student's total vocabulary size, which is important for reading and general educational development" (Jamestown Publishers, Catalog for 1982).

As is stated in the examiner's manual, the estimation of vocabulary size based on this test does not represent the total number of words an individual knows, but rather, the total number of Basic Words, as they have been defined in Dupuy (1974). Dupuy did succeed in giving an explicit, operational definition to the construct "Basic Word." It is

very questionable, however, whether this construct can be given the interpretation that the name "Basic Word" suggests. Our results indicate that Dupuy's estimate of 12,300 basic words in English is a gross underestimate of the number of distinct vocabulary items in the language. Our figure of 88,533 distinct word families is larger than Dupuy's by a factor of seven. If we define total number of words in terms of items that must be learned individually--counting homographs and other distinct meanings, abbreviations, etc., as separate words--the number of words in printed school English may be as high as 110,000. Thus, the true vocabulary size of an individual could be more than seven times greater than what is indicated by his or her performance on the Basic Word Vocabulary Test.

Of course, it is not possible to get an accurate revised measure of vocabulary size simply by multiplying scores on Dupuy's test by seven. The items in the test, although they may be a representative sample of Basic Words as defined in Dupuy, do not necessarily constitute a representative sample of basic words in any other sense. In addition, while our estimate of the total number of distinct words in English is seven times greater than Dupuy's, a quite different relationship may hold between specific subsets of these words. For example, the number of items among our distinct word families that a third grader would be likely to know may not be seven times as great as the number of Dupuy's Basic words that fall into this same category.[5] Still, it is possible to conclude that the Basic Word Vocabulary Test underestimates vocabulary size by an order of magnitude.

Programs of Vocabulary Instruction

Our results indicate that the number of words that students encounter in reading is very large, and the results strongly suggest that children's vocabularies are larger than some recent investigators have supposed. Advocates of direct vocabulary instruction have leaned heavily on the assumption that the number of distinct words in school English is small, that unaided year to year growth in vocabulary is modest, and that the total number of word meanings known by a typical child at any age is not large. Notably, Becker, Dixon and Anderson-Inman (1980), accepting Dupuy's estimate that the average high school senior knows approximately 7,800 words, have attempted to lay out a program of systematic instruction for a core vocabulary of 8,000 words.

Our findings suggest that high school students may actually know far more words, perhaps somewhere between 25,000 and 50,000, or even more. Dupuy (1974) estimates that third graders know only 2,000 words, but estimates by others are higher. Cuff (1930) places third grade vocabularies at around 7,425 words, and M. K. Smith (1941), using vocabulary tests based on Seashore and Eckerson (1940), set the figure at 25,000 basic words. It is quite possible, then, that the average third grader already knows 8,000 words.

A program of systematic instruction for a core vocabulary of 8,000 words is not necessarily a bad idea. As Table 10 shows, if 8,000 words were correctly chosen, they could cover all distinct word families found among words that occur at least once in a million words of text. But the

theoretical foundation of this program--taking Dupuy's Basic Words as a benchmark for the number of items to be learned--is questionable.

There is reason to worry that Becker,Dixon, and Anderson-Inman did not find the right set of 8,000 words, and, furthermore, that they made unreasonable assumptions about semantic relatedness. They culled their set of 8,000 words from a list of 26,000 based on the Thorndike and Lorge (1944) Teacher's Word Book of 30,000 words, with some adjustments to bring the list up to date. The list of 26,000 "object words" was collapsed to 8,000 "root words," where a root word was defined as "the smallest word from with the other words can be semantically derived....In designating a root word for any given object word a search was made for the smallest word within the object word that contains the core meaning of the object word" (emphasis in the original). The assignment of root words was frequently the same as in the present analysis; for example, the root word of helpless was help. However, in our judgement, Becker and his associates often stretched the criteria of semantic and morphological relatedness beyond reason. For example, all of the following words were assigned the root word judge on the basis of their semantic relatedness: juror, juridicial, jurisdiction, jurisprudence, jury, judicious, judicature, prejudice, prejudicial, unprejudiced, judicial, judiciary, judge, and judgement.

The problem with this grouping is the assumption that direct instruction on the root words and on affixes would automatically result in a child knowing the meanings of the whole set of words. Becker,

Dixon, and Anderson-Inman (1980, p.7) admit that "providing systematic instruction for even 8,000 root words is a monumental undertaking." We consider it even more monumental for a student, having been taught only the meaning of judge, to be able to identify what words were in fact related to it, and then to figure out their meanings. How could a child, encountering words such as Judaic, judicious, judo, juggernaut, juggle, jugular, Julian, junta, and jury for the first time in text, know which were historically related to judge? Furthermore, the most important part of the meaning of a word such as jury is not what it has in common with the root word judge (this much of its meaning would probably be pretty obvious from the context), so much as how it differs from it. Furthermore, since the root words were usually chosen to be one of the more frequent members of a set of related words, it may well be that children already know many or most of the 8,000 root words, and that it is the "derived" words such as judicial, jury, and judiciary, rather than root words like judge, for which they really need instruction.

Of course, many of the derived words were in fact transparently related to their root words. But because no distinctions were made among different degrees of relatedness or different types of relatedness, Becker and his colleagues underestimate the number of words that are functionally distinct as far as vocabulary learning is concerned.

Beck, McCaslin, and McKeown (1980) have formulated an intensive program of vocabulary instruction which has as a major aim increasing student's reading comprehension. One motivation for their program was

that several previous experimental studies have failed to produce significant increases in reading comprehension via vocabulary instruction (e.g., Jenkins, Pany, & Schreck, 1978). Beck and her associates hypothesize that vocabulary instruction can facilitate reading comprehension only if the words are learned thoroughly--to the point where the word's meaning can be accessed quickly or automatically, and where a fairly rich network of semantic connections between that word and others has been developed. Because of this, their program involved repeated exposure to words. Children in their study were exposed to each word 10-18 times in a variety of tasks. There was also a subset of words in their study which were repeated 26-40 times, to see if the additional repetition would result in even greater learning.

Results from an application of this program in a fourth grade classroom are described in detail in Beck, Perfetti and McKeown (in press). Even with the intensive instruction and repetition, children learned 77.6% of the words that were repeated 10-18 times, and 86.5% of the words repeated 26-40 times. So it does not appear that the program was unnecessarily repetitive.

How much ground did the program cover? Just 104 words were taught over a five month period, with one half hour per day devoted exclusively to this vocabulary program. At this rate, 208 words could be covered in a school year. If the program were streamlined by having all words repeated only 10-18 times (that is, dropping the extra repetition of the special subset of words), one might be able to cover a little over 400

words per year. Note that Becker, Dixon, and Anderson-Inman's program to cover 8,000 words in 10 years would have to progress at twice this rate, either by spending more total time on vocabulary, or less time on each word.

How does this compare with the amount of vocabulary that students encounter in school? According to our rough estimates, a child might easily come across a thousand or more totally new word families each year in his or her reading; for an active reader in the upper grades, the figure would certainly be higher. Thus, the program of vocabulary instruction suggested by Beck and her asssociates could not hope to cover half of the new words children actually encounter in their school reading. And the total number of words covered by such a program in ten years of school--at most around 5,000 words--would apparently constitute only a small fraction of the reading vocabulary of a fairly good reader.

According to Beck, McCaslin and McKeown (1980, p. 8) it takes "an extended series of fairly intensive exposures [to a word]...before it can be quickly accessed and applied in appropriate contexts." It may well be, of course, that automaticity of access is the key factor in the relationship of word knowledge to reading comprehension; but the puzzle that must be solved by those who propose to produce automaticity using word drills is how to do it in the available time, not just for four or five thousand words, but thousands or even tens of thousands of less frequent ones.

The schools have never had programs of vocabulary instruction as extensive as that proposed by Becker or as intensive as that proposed by Beck. The question that naturally arises is, up to now, how have readers acquired their vocabulary knowledge? Our answer to this question appears in the final section of this paper.

Generalization to Non-Instructed Words

A basic implication of our study is that, because of the sheer volume of vocabulary that students will encounter in reading, any approach to vocabulary instruction must include some methods or activities that will increase children's ability to learn words on their own. Any attempt to do this would be based on one or more of three possible emphases: Motivation, inferring word meanings from word parts (morphology), and inferring word meanings from context.

There is basically no experimental literature that could confirm the success of any of these in facilitating children's learning of words on their own. We can at least speculate, though, on the implications of our findings as to the effectiveness of such approaches.

With respect to motivation, it is no doubt an important factor. For all we know, it may be as important as any other aspect of vocabulary instruction. To quote from Petty, Herold and Stoll (1968),

> [M]any researchers considering vocabulary development pass over motivation without mention. No classroom teacher genuinely attempting to teach vocabulary makes that mistake....[T]eachers reporting on favorite techniques begin with discussions of how student interest in word study was created (p. 19).

Beck's program does include a strong motivational component. For instance, some of the learning activities took the form of competitive games, and there were incentives for children to report instances of instructed words they found outside the classroom. Attention to motivational factors did seem to contribute to the overall success of the instruction. Beck and her colleagues feel it may be a reason for the apparent increase in the experimental children's performance on tests of words not covered in the instruction. However, further research will be necessary to determine whether this effect was really a generalized increase in word learning, the result of improved vocabulary test taking skills, or an artifact of experimental design.[6]

Morphology and Vocabulary Instruction

Our findings suggest an important role of morphology in the learning of vocabulary. Semantically transparent derived words are relatively rare among the most frequent words, but constitute an increasinly greater proportion of the vocabulary as one goes towards the lower end of the frequency continuum.

For this reason, frequency cannot be the only criterion by which words are chosen to be included in vocabulary instruction. If the students only encountered words of fairly high frequency, there would be little opportunity to learn the productive word-formation processes in

English that constitute the key to understanding the bulk of lower-frequency words.

The introduction of new words should be determined by family relationships as well as by frequency. For example, drama and dramatic are fairly frequent words (with Us of 11 and 18, respectively), but the derivative forms are fairly rare in printed texts, e.g., dramatist (U = .02), dramatize (U = .40), and dramatization (U = .50). Teaching words together as a family has a number of advantages. First, if the most frequent words in the family are already known, this procedure builds a bridge from familiar to new. In any case, once the meanings of drama were instructed, the meanings of the derivatives could be covered with little additional effort. What additional time is devoted to the derivatives would also function to reinforce the learning of the base word as well.

Another benefit of teaching words in families would be to call the students' attention to the word-formation processes that relate the different memebers of the family, so that they would be more likely to take advantage of such relationships on their own. In addition, covering a family of words would familiarize students with the types of changes in meaning that often occur between related words, thus preparing them to deal with cases in which the semantic relationships among morphologically related words are not so transparent.

It should be remembered, however, that our definition of word family is based on relationships among existing words in English, not on

historical roots, and on semantic relationships that are transparent enough for students to perceive on their own. We remain highly skeptical of approaches to vocabulary that proceed on an etymological or historical approach to word meanings, approaches which feign that words such as dialect, collect, and intellect have some basic meaning in common. There may be some perceptual or mnemonic value to analysing words into historically-based components, but this remains to be established. Shepherd (1974) found that knowledge of Latin roots (e.g., -ceive, lect) is not strongly related to the knowledge of the meanings of words containing such roots (e.g., receive, collect), whereas knowledge of stems which themselves are English words (e.g., sane) is strongly related to knowledge of the meanings of related derived forms (e.g., sanity). The type of relatedness among words analysed in the present study, along with its associated implications for instruction, is not to be confused with the etymological or historical approach adopted by some.

Learning Word Meanings from Context

That word meanings are learned from context is an inescapable fact. Many ninth graders, even more high school seniors, and almost all educated adults would be able to read with comprehension through any school materials for grades three through nine with a high level of comprehension. This presumably requires knowing a large proportion of 88,500 distinct word families. These words could not be acquired from direct instruction nor from looking them up in a dictionary. There is only one other possible source of knowledge: Inference based on context.

Thus, logic forces the conclusion that successful readers must learn large numbers of words from context, in most cases on the basis of only a few encounters.

It is hard to conceive how a word such as if, for example, could be learned in any other way than from verbal context. Pointing to something in the world that corresponds to the concept of hypotheticality would be difficult to say the least, and any child old enough to understand a non-circular definition of if is surely already able to use the word fluently.

Good readers may acquire large vocabulries exactly because they are better at inferring word meanings from context. One indication of this is the fact that a cloze test is a satisfactory measure of reading ability. While a cloze test is taken as indicating overall reading ability, the skill it measures most directly is the ability to use contextual information to supply the meanings of words missing from text--a task analogous to that of identifying the meanings of unknown words in context.

Knowledge of morphological relatedness among words proably contributes importantly to learning word meanings from context.[7] Our findings here show that a large number of infrequent words are transparent derivatives of other words, in many cases of words the student is likely to know already. While context often is not sufficient to determine the meaning of an unfamiliar word, it may provide enough information to permit a guess at the appropriate meaning of a word whose

semantic content is partially determined by its morphology. A child who knows the meaning of drama and the function of the suffix -ist will need only minimal help from context to determine the meaning of dramatist. A hypothesis that should be explored in future research is that joint utilization of contextual and morphological information is a strategy employed by children who develop large vocabularies.

We hypothesize that the principal engine driving vocabulary growth is volume of experience with language. Oral language experience is important, of course, particularly for the young child, but we judge that beinning in about the third grade the major determinant is amount of free reading. It is a surprising fact that there are no satisfactory estimates of the number of words read per year by children of different ages. Earlier we guessed that the least able and motivated children in the middle grades might read 100,000 words a year while average children at this level might read 1,000,000. The figure for the voracious middle grade reader might be 10,000,000 or even as high as 50,000,000. If these guesses are anywhere near the mark, there are staggering individual differences in volume of language experience, and, therefore, opportunity to learn new words. Notice also that variation of this magnitude could readily explain differences between good and poor readers in automaticity of word access.

The only thing problematical about the "rapid learning from context" theory is that experimental studies generally have seemed to show that children do not learn word meanings very well from context. For

instance, Jenkins, Pany and Schreck (1978) found that exposure to words in context produced little increase in knowledge of their meanings, and no measurable increase in the comprehension of text containing those words. Two factors may account for this finding. First, there is reason to doubt whether the contexts used in this experiment were really suitable for learning the meanings of the new words. Second, as Jenkins, Pany, and Schreck suggest, it may be that readers can encounter a substantial number of unfamiliar words in a text and still comprehend it fairly well, especially if they have some acquaintance with the general subject matter. Whatever the explanation, the failure to find experimental evidence for contextual learning of word meanings ought to be regarded as a conundrum for experimentalists rather than the basis for educational policy.

References

The American Heritage school dictionary. Boston: Houghton Mifflin, 1977.

Anderson, R. C. & Freebody, P. Vocabulary knowledge. In J. T. Guthrie (Ed.), Comprehension and teaching: Research reviews. Newark, Del.: International Reading Association, 1981.

Anderson, R. C. & Freebody, P. Reading comprehension and the assessment and acquisition of word knowledge. In B. Hutson (Ed.), Advances in reading/language research, a research annual. Greenwich, Conn.: JAI Press, in press.

Aronoff, M. Word formation in generative grammar. Cambridge, Mass.: M.I.T. Press, 1976.

Beck, I., McCaslin, E., & McKeown, M. The rationale and design of a program to teach vocabulary to fourth-grade students. Pittsburgh: University of Pittsburgh, Learning Research and Development Center, 1980.

Beck, I., Perfetti, C., & McKeown, M. The effects of long-term vocabulary instruction on lexical access and reading comprehension. Journal of Educational Psychology, in press.

Becker, W., Dixon R. & Anderson-Inman, L. Morphographic and root word analysis of 26,000 high frequency words (Tech. Rep. 1980-1). Eugene, Ore.: University of Oregon Follow Through Project, April 1980.

Berko, J. The child's learning of English morphology. Word, 1958, 14, 150-177.

Campbell, D. T. & Boruch, R. F. How regression artifacts can mistakenly make compensatory education look harmful. In C. A. Bennitt and A. A. Lumsdaine (Eds.), Evaluation and Experiment: Some critical issues in assessing social programs. New York: Academic Press, 1975.

Carroll, J. B. Measurement properties of subjective magnitude estimates of word frequency. Journal of Verbal Learning and Verbal Behavior, 1971, 10, 135-142.

Carroll, J. B., Davies, P., & Richman, B. The American Heritage word frequency book. Boston: Houghton Mifflin, 1971.

Cuff, N. B. Vocabulary Tests. Journal of Educational Psychology, 1930, 21, 212-220.

Deighton, L. C. Vocabulary development in the classroom. New York: Bureau of Publications, Teachers College, Columbia University, 1959.

Dupuy H. P. The rationale, development and standardization of a basic word vocabulary test. Washington, D.C.: U.S. Government Printing Office, 1974. (DHEW Publication No. HRA 74-1334)

Dupuy, H. P. Basic Word Vocabulary Test. Highland Park, N.J.: Dreier Educational Systems, 1975.

Funk and Wagnalls New Standard Dicationary of the English Language. New York: Funk & Wagnalls Co., 2 Vol. unabridged edition, 1937.

Funk and Wagnalls New Standard Dictionary of the English Language. New York. Funk and Wagnalls Co., 1965.

Harwood, F. W., & Wright, A. M. Statistical study of English word formation. Language, 1956, 32, 260-273.

Jamestown Publishers Catalog for 1982. Providence, R.I.: Jamestown Publishers, 1982.

Jenkins, J. R., Pany, O., & Schreck, J. Vocabulary and reading comprehension: Instructional effects (Tech. Rep. No. 100). Urbana: University of Illinois, Center for the Studey of Reading, August 1978. (ERIC Document Reproduction Service No. ED 160 999)

Lorge, I., & Chall, J. Estimating the size of vocabularies of children and adults: An analysis of methodological issues. Journal of Experimental Education, 1963, 32, 147-157.

Petty, W. T., Herold, C. P. & Stoll, E. The state of knowledge about the teaching of vocabulary. Urbana: National Council of Teachers of English, 1968.

Random House dictionary of the English language. New York: Random House, 1966.

Rhode, M., & Cronnell, B. Compilation of a communication skills lexicon coded with linguistic information (Tech. Rep. No. 58). Los Alamitos, Calif.: SWRL Educational Research and Development, November 1977.

Seashore, R. H., & Eckerson, L. D. The measurement of individual differences in general English vocabularies. Journal of Educational Psychology, 1940, 31, 14-38.

Shepherd, J. F. Research on the relationship between meanings of morphemes and the meanings of derivatives. In P. L. Nacke (Ed.) 23rd N.R.C. Yearbook. Clemson, South Carolina: National Reading Conference, 1974, 115-119.

Smith, M. K. Measurement of the size of general English vocabulary through the elementary grades and high school. General Psychological Monographs, 1941, 24, 311-345.

Stauffer, R. G. A study of prefixes in the Thorndike list to establish a list of prefixes that should be taught in the elementary school. Journal of Educational Research, 1942, 35, 453-458.

Thorndike, E. L. The teacher's word book of 10,000 words. New York: Teachers College Press, 1921.

Thorndike, E. L. The teaching of English suffixes. New York: Teachers College Press, 1941.

Thorndike, E. L., & Lorge, I. The teacher's word book of 30,000 words. New York: Teachers College Press, 1944.

Webster's third new international dictionary (unabridged). Springfield, Mass.: Merriam Co., 1961.

The world book dictionary. Chicago: Chicago Field Enterprises Educational Corp., 1969.

APPENDIX A

Categories of Relationships Among Words

Morphologically Basic Words

This category includes any words which cannot be described as related to some more basic word via some productive or semi-productive word formation process. First of all, this means any monomorphemic words, e.g., add, foil, or wind. It also includes words that might be considered multimorphemic in a historical sense, but which do not seem analysable in terms of the word-formation processes of modern English.

Operationally, this category is also the "none of the above" category, that is, the classification of words which do not fall into the other relationship categories in our coding system. However, if we have bent criteria, it has normally been in the direction of coding an item in some other relationship category. For example, the category of "idiosyncratic morphological relationships" was used to categorize relationships (e.g., between knowledge and know) which would not be considered productive word formation processes of modern English.

This category also includes those items which are morphologically basic as far as the American Heritage Intermediate Corpus is concerned. For example, the word imposters occurs in the corpus, but not the singular imposter. Since no other words related to this item occur in the corpus either, it was coded in the category "morphologically basic with respect to this corpus." Items in this category were included with the category "morphologically basic words" for the purpose of counting

types of relatedness, although they are also distinguished from the truly basic words by a special flag.

Simple Capitalization

This category includes all items in the corpus which differ from some other existing item only with respect to capitalization. For example, Teacher differs from teacher only in the capitalization of the initial letter. This category is called simple capitalization in that it does not include cases of capitalization homographic with a proper name, e.g., Jets or Earl. Such items are included in the category "Capitalizations homographic with proper names," discussed below.

Alternate Spellings

This category includes those items which differ from some other item only with respect to spelling. For example, cart-horse is treated as a spelling varient of carthorse. In many cases, this category was used for misspellings which occurred in the corpus.

Alternate Pronunciations

This category was used for items spelled in nonstandard ways to indicate pronunciation, for example, fishin', or crrrack.

Alternate Form of Word

This category was used for alternate forms of words such as soya and soy, hurray and hurrah, or britches and breeches, where the difference in spelling reflectes a difference in pronunciation, but one which involves

the phonemic form of the word. In other words, this category covers minor differences in lexical form, whereas the category "Alternate Pronunciations" covers differences which might be thought of as resulting from low-level phonetic rules.

Alternate Forms with S

This category is a special case of the previous one. It includes those minor variations in lexical form which consist of the presence or absence of a final s, as in toward and towards or amidship and amidships. For lack of a better category, the pair amid and amidst is also categorized here.

Regular Inflections

This category includes all items related to their immediate ancestors by regular inflection--that is, items which differed from other items only by the endings s (es), ed, ing, 's, and s'. Since the WFB provides no context, it was not possible to distinguish between contractions (John's = John is) and possessives. Therefore, in cases where a form ending in could be interpreted as a possessive, it was included among the regular inflections.

In the coding system there was a distinction made between regular inflections (i.e. plurals, possessives, past tenses or past participles, and third person singulars of verbs) and instances where ed or ing result in words with distinct syntactic and perhaps also semantic properties, as in the case of spelling, planking, crowded, and elevated. This

distinction, however, was often difficult to make. There are some cases, such as dress/dressing, where there are sustantial semantic shifts between the two words; about 20 such items were found among the words coded. In other cases, the semantic differences are a little less pronounced, as in the case of spell/spelling. The semantic aspect of the coding system will have captured the important differences between these types of relationships. For the purpose of the overall counting, it was decided to lump together all regular inflections, including items such as spelling or dressing. The semantic codes can be used to distinguish such cases when necessary.

The following categories were coded as distinct from regular inflections;:

a) Semantically irregular plurals such as top/tops, air/airs, and premise/premises.

b) "Scientific" plurals such as genetics and genitals.

c) Incorrect regular inflections such as knowed.

d) Alternate forms of words with s, such as skyward/skywards.

Only 21 of the 7260 items coded fell into these last four categories.

Irregular Inflections

This category includes irregular plurals of nouns (mouse/mice), irregular past tenses and participles of verbs (tear/tore/torn), some Latin plurals (larva/larvae), and also suppletive forms such as I, me, mine. Also included in this category are suppletive forms of the verb to be, for example, is, are, was, were, been. Included as well in this category were relationships such as our/ours, and my/mine.

As with regular inflections, there was a separate coding category for irregular inflections that resulted in distinct words with different syntactic (and sometimes semantic) properties. For example, known functions as an adjective (a known criminal), as well as a past participle (he should have known the answer). As in the case of regular inflections, this distinction was sometimes difficult to make, and was not incorporated into the counts presented here; both types of irregular inflections were lumped together. Cases where there is a distinct semantic difference between the two synntactic uses of the word can be identified in terms of the semantic coding distinctions to be discussed below.

Regular Comparatives and Superlatives

This category includes forms such as faster, slower, quickest, and highest.

Irregular Comparatives and Superlatives

This category includes forms such as better, best, and worst.

Suffixation

Target items related to their immediate ancestor by suffixation were divided into four categories: First, what could be called "normal suffixation." This is best defined in terms of the three remaining categories which can be distnguished from it. The second category might be called "suffix replacement." This category is used for those cases in

which the target word has a different suffix than its immediate ancestor. This will necessarily be the case when the stem does not occur in English without an affix. For example, the immediate ancestor of aggressive is aggression (cf. Aronoff, 1976). Similarly, the immediate ancestor of enthusiastic is enthusiasm. The same holds for pairs such as chloride/chlorine, or stenographer/stenography. It was also decided to treat pairs such as fragrance/fragrant and omnipotence/omnipotent in this fashion.

A third subcategory of suffixation includes those cases where the addition of a suffix is accompanied by unpredictable changes in the form of the stem: for example implication/imply, apathetic/apathy, negligent/neglect, or sensuous/sense. A fourth subcategory of suffixation was used for those cases in which it seemed proper to analyse a word into a stem + suffix, even when the stem itself was not an English word. For example, nomin + al, cruci + fy. Only three cases of the 7260 items coded were put into this category.

Prefixation

Target items related to their immediate ancestors by prefixation were similarly divided into four categories: Examples of "prefix replacement" are pairs such as decrease/increase and descend/ascend. Cases where prefixation involved unpredictable changes in the form of the stem included impoverish/poverty and mishap/happen.

No cases were analysed as prefix + bound stem. This would be done only where there was some justification for assigning some specific semantic content to the stem; this cannot be done in cases such as deceive, perceive, or receive (cf. Shepherd, 1974).

Compounds

Compounds were coded into seven subcategories.

First, there are regular compounds--those which do not fall into any of the following special categories. Second are hyphenated compounds which do not meet criteria for any of the following special categories. The difference between these first two categories is simply spelling. It is not clear whether hyphens are used in compounds with any regularity or consistency, but it seemed best to code the two types as distinct, since the categories can always be collapsed afterwards. We have not made any use of the distinctions among compound types in the analyses presented here.

Third are hyphenated compounds with the internal structure of phrases or sentences--for example: doctor-to-be, fission-fusion-fission, twenty-year-old, or live-and-let-live.

A fourth category of compounds are contractions, such as can't, daddy'll, nobody'd, and would've.

A fifth category of compounds was used for cases where the component parts of the compound were not free stems in English, but could be assigned a specific semantic value; for example omnipresent, cartography, theology, or automobile.

A sixth category of compounds was used for those involving an adverbial particle: wind-up, burnout, hookup, and tie-in. A final category was used for compounds such as cranberry or chamberlain where one element was clearly a meaningful unit in English, but the other was not a word in English, nor could it easily be assigned any specific semantic value.

Truncations

This category was used for the relationship between such pairs as rhinocerous/rhino, racoon/ˆcoon/ and gentleman/gent. These cases were distinguished from abbreviations, such as Mich for Michigan.

Idiosyncratic Morphological Relationships

This category was used for items which seemed to show a definite morphological relationship with some immediate ancestor, yet which did not seem to belong in the other categories. Often, this involved a difference in form that could be thought of as a suffix, but was not productive at all in English. For example, there were pairs suc as: largesse/large, prophesy/prophecy, musicale/musical, planetarium/planet, or knowledge/know.

Ambiguities

The WFB was collected by computer, with "word" being defined as a string of characters bounded right and left by spaces. This definition treats as distinct words any graphically distinct types, no matter how trivial the difference. It also lumps together any graphically identical

types, no matter how semantically diverse--all the different meanings of bat, or mean, or bear. It would not be possible, and hence it was not our intention, to disambiguate the items in this corpus. We have dealt with one specific type of ambiguity, however: what could be called morphological ambiguity, or ambiguity of relationship category. That is, we have tried to represent ambiguity when it involved a word being analysable into two or more of our categories of relatedness. A word such as bat, for example, however many meanings it may have, falls into only one category of relationship; it is a morphologically basic word. The word bats, similarly, may have a number of meanings, but its relationship type is unambiguous: it is a regular inflection of bat. The word felt, on the other hand, is ambiguous in terms of its morphological relationships. On one hand, it is an irregular past tense of the verb feel (which may of course have any number of meanings). On the other hand, it is a morphologically basic word as well.

A word such as felt was coded as being related to two (or more, when necessary) items, felt1 and felt2. These latter items, by definition unambiguous with respect to their morphological relationships, were then further analysed as any other items in the list would be.

APPENDIX B

Target Word - Immediate Ancestor Pairs

Illustrating SEM 0

TARGET WORD	IMMEDIATE ANCESTOR
senselessly	senseless
sensibly	sensible
chlorination	chlorinate
cleverly	clever
cleverness	clever
daintiness	dainty
decentralization	decentralize
desecration	desecrate
desegregation	desegregate

Target Word - Immediate Ancestor Pairs

Illustrating SEM 1

TARGET WORD	IMMEDIATE ANCESTOR
elfin	elf
geneticist	genetic
misrepresent	represent
fragmentary	fragment
litigant	litigate
sunbonnet	sun
enthusiast	enthusiasm
washcloth	wash
collectively	collective
anywhere	any
crowded	crowd
various	vary
lower-class	lower
wily	wile
wind-twisted	wind
yummy	yum
Botanic	botany

Target Word - Immediate Ancestor Pairs

Illustrating SEM 2

TARGET WORD	IMMEDIATE ANCESTOR
therapeutic	therapy
gunnery	gun
gunner	gun
foglights	fog
uncountables	uncountable
cow-hand	cow
mainly	main
additional	addition
knowledge	know
once	one
everyday	every
sky-high	sky
space-sick	space
stringy	string
sun-suit	sun
sunburn	sun
theorist	theory

Target Word - Immediate Ancestor Pairs

Illustrating SEM 3

TARGET WORD	IMMEDIATE ANCESTOR
password	pass
handspring	hand
collarbone	collar
airfoil	air
bloodshot	blood
sensor	sense
skydiver	sky
tweeter	tweet
visualize	visual
washroom	wash
apeak	peak
Sunday-school	Sunday
hookworm	hook
inlay	lay
mishap	happen
moonship	moon
noblesse	noble
ominous	omen
passenger-miles	passenger
pasteurize	Pasteur
percentile	percent
planetarium	planet
broadax	broad
chloride	chlorine
collinear	linear
conclusive	conclusion
doctorate	doctor
doctrinaire	doctrine
elevator	elevate
fishwheel	fish

Target Word - Immediate Ancestor Pairs

Illustrating SEM 4

TARGET WORD	IMMEDIATE ANCESTOR
crowbait	crow
saucepan	sauce
fender	fend
vitality	vital
high-school	high
saucer	sauce
artificial	artifice
apartment	apart
colleague	league
condescend	descend
go-getter	go
impregnable	impregnate
impressionable	impression
moonstruck	moon
negligible	neglect

Target Word - Immediate Ancestor Pairs

Illustrating SEM 5

TARGET WORD	IMMEDIATE ANCESTOR
dog-days	dog
Burma-Shave	Burma
prefix	fix
peppermint	pepper
shiftless	shift
misgive	give
poochie-pies	pies
crowbar	crow
foxtrot	fox
livelong	live

APPENDIX C

Types of Words in the Corpus

One issue in determining vocabulary size is deciding what types of words to count, i.e. whether to include proper names, abbreviations, numbers, and so on. We used the following set of categories to classify the items in the WFB:

Proper Names

This category was used primarily for names of specific individuals (whether historical or fictional), and for names of geographic places. Words directly derived from such proper names (e.g. American, Burmese, British-controlled) were also included. Coded in this category as well were days of the week, months, and names of companies and organizations (as well as abbreviations of such names, e.g., AMF, AKC). Capitalization was taken as evidence, but was not used as a criterial factor.

Items Homographic with Proper Names

In many cases, a capitalized word could be taken either as a proper name (or part of a proper name), or else a common noun capitalized for some other reason: e.g., Dodge, Drew, Cook, Dipper, Campfire, Earl, Hood, Jets. (Because of the way the WFB was collected and keypunched, many common nouns occur both in capitalized and uncapitalized form.) The category of items homographic with proper names was grouped together with the category "Proper Names" for the purpose of the analyses reported here. This is because they allow interpretation as proper names, and

their uncapitalized versions have been already included in determining the number of non-proper names.

Numbers and Formulae

This category includes types such as AOG, MCVII, NXN, R5, 1089, and 85%.

Compounds or Derivatives Based on Numbers

This category includes types such as 32nd, 106-ton, 17th-century, and 82-degree.

Abbreviations

Only twelve items of the 7260 coded fell into this cagtegory: They were fps, Md, NW, PX, Rw, RW, TD, Te, MD'S, Doctr, and Ave. Dictionaries were used to distinguish abbreviations from formulae. The subject categories in the WFB also helped determine the proper interpretation of some items; for example, if the type AOG occurred only in Mathematics, it would seem to be best interpreted as a formula (probably the name of an angle), rather than as the name of some organization.

Foreign Words

This category included words recognizable as belonging to languages other than English, were were not found in the reference dictionaries used; for example: ponere, daeghwamlican, Romani, les, las, Irae, decem, and noire.

Nonwords

In this category were listed items which were not found in the reference dictionaries used (including Webster's Third New International, unabridged), and which could not be assigned to any of the other coding categories discussed here. Some of the items found in this category are clearly onomatopoetic: putt-putt-putt or wh-i-s-s-s-t. Others may be deliberate coinages, such as yugit, clicket, or pickie. Still others may be noncapitalized versions of unfamiliar proper names (maribou, faeger), or misspellings of other words. The total number of items in this category (147) is small enough so that reclassification of some of them would not have much effect on the overall distribution of types in our analyses.

"WFB Errors"

A final category was used for 6 items which were erroneously repeated in both the book and tape versions of the WFB.

Footnotes

The research reported herein was supported by the National Institute of Education under Contract No. US-NIE-C-400-76-0116.

[1]It should be noted that the addition of such items to the list does increase the overall size of the list, but does not inflate the number of items in any given category. To illustrate this, consider a hypothetical list consisting only of the words abatement, abates, abated, and after. As it stands, the total length of the list is four items; in terms of relationship categories, there would be one instance of suffixation, two instances of regular inflection, and one basic word. Our goal, however, is to define the count so as to have it reflect the number of word families in a corpus, for any given definition of word family that can be constructed in terms of our coding system. For example, assume that we want to know the number of distinct word families in this hypothetical corpus for a child who understands regular inflections, but who has not yet internalized any rules of suffixation. For such a child, there would be three distinct word families in this corpus: One containing after, one containing abates and abated, and one containing abatement. (We had assumed that the child at this point did not recognize the connection between abatement and abate.) If we add the missing ancestor abate to the list, to arrive at the number of distinct word families, we simply take the number of basic words, plus the number of items in any relationship type not yet mastered by the child at the level of linguistic development in question. In this case, the corpus would contain abate (the missing

ancestor of abates and abated), abates, abated, abatement, and after. That is, two basic words, two regular inflections, and one instance of suffixation. If we want to know how many word families are in the corpus for a child who has internalized the rules of regular inflection, but not those of suffixation, we arrive at the count of three. For a child who has also mastered suffixation, there are only two distinct word families in this corpus.

Thus, the addition of "missing ancestors" to the list does increase the overall number of items, but it does not distort the count of items in any given relationship category. The same holds for items added to disambiguate morphologically ambiguous target words. Consider a hypothetical corpus consisting of the following items: feel, felt, go, went and after. We would want to say that there are four morphologically basic words, feel, go, after, and the noun felt. We would also want to say that the list contained two irregular inflections: went and felt. Thus, a morphologically ambiguous word like felt should be counted in each of the categories to which it belongs.

Thus, tabulations of the number of items in various relationship categories will include added entries which are disambiguations and missing ancestors, in determining the composition of the sample and the corpus.

There were also certain items added to the list during the coding process which were not included in tabulation of relationship types. For example, compounds were given a separate entry for each component part.

This was because the relationship between farmhand and farm, for example, might be quite different than the relationship between farmhand and hand. The first relationship is semantically transparent; the second involves a secondary meaning of hand related to the more primary meaning by a metaphor (metonomy might be the more accurate term in this case) which might not be immediately transparent to an elementary school child. In any case, for each compound, additional items were added to express the relationship of the compound to each of its component parts. This added items were not, however, counted in the tabulation of the number of items in any given relationship category.

In the tabulation of compounds for different levels of semantic transparency the two codes for each compound were collapsed, and the compound was assigned the degree of semantic transparency associated with the least transparent of its members. This reflects the assumption that the difficulty of learning a new compound such as farmhand is determined largely by the difficulty of learning the least semantically transparent of its component parts.

[2]The values in our estimates for the population of words in printed school English were calculated as follows: First, the items in our sample were ordered by frequency, and divided into seven strata containing equal numbers of items, each representing a band of frequencies. From Table B-8 in the WFB (p. xxxvi), the number of words in printed school English within each frequency band was determined. A weighting factor was assigned to each stratum representing the ratio of

the number of words in the population within that frequency band to the number of words in the corresponding stratum in our sample.

The size of the WFB, even as large as it is, creates an artificial "floor" for the reported frequencies. That is, any word, however low its "true" probability or frequency, if it occurs in the corpus at all, will be assigned a certain minimum frequency value. The U-values (estimated frequency per million) of the 35,079 hapax legomena in the corpus were adjusted according to the amount of text from the subject categories in which they occurred. The result of this was that the second from the lowest frequency stratum in our sample had an artificially small frequency range (in terms of reported frequencies), and hence an unrealistically low weighting factor in the initial estimate. This was corrected by plotting the final weighting factors on a smooth, essentially exponential curve determined by the value of the other weighting factors and by the constraints on the value of the sum of all weighting factors.

The actual weighting factors had the following values, expressed in terms of how many words in the population a single word in each stratum of our 7260-word sample would represent.

STRATUM	FREQUENCY RANGE		WEIGHT
	LOWER U	UPPER U	
1	.0004	.0109	314.80
2	.0109	.0150	121.17
3	.0150	.0457	64.78

4	.0457	.1176	38.39
5	.1176	.4071	23.39
6	.4071	2.0430	13.80
7	2.0430	7456.8281	12.72

The weights given are those relating our sample to the population; the relationships between the WFB and the population could be represented by dividing those weights by 11.9478.

We also wanted to determine the extent to which the choice of weighting factors influenced our final estimates of vocabulary size. Therefore, we tried calculating estimates for the total population on the basis of a number of sets of weighting factors--the original estimates, our adjusted smooth exponential curve, and a number of exponential functions which in effect defined the extreme values of functions that could be drawn through the points determined from the tables in the WFB.

Our final weighting function gave us an estimate of 45,453 morphologically basic words in the population. The other sets of weighting factors gave estimates ranging between 45,285 and 47,418 morphologically basic words, a range of only 2,133. Thus, any reasonable variation in the weighting factors would lead to only very small differences in the values of our final estimates. Even for those categories more skewed in terms of frequency than were the morphologically basic words, the estimates based on the different sets of weighting factors were very close.

We also calculated estimates for the population by assigning weighting factors to words individually on the basis of the function

$$W = 11.9478/(1 - (1 - p)^n)$$

where 11.9478 is the number of words in the WFB divided by the number of words in our sample, and p is the probability of a word, that is, U/1,000,000. The expression $(1 - (1 - p)^n)$ is the likelihood of a word with probability p occurring in a corpus of n running words; hence it is also the proportion of words with probability p that should occur at least once in a corpus of n words. This formula gave us essentially the same results as our earlier calculations.

Note that items added to the original sample in the coding process--missing ancestors and disambiguations--were not included in the process of estimating the composition of the population. The procedures for extrapolating from the sample to the population already account for words that do not occur in the WFB, so to include items added to our sample in these estimates would have amounted to counting them twice.

Morphologically ambiguous items were also not included in our projections for the population, because there was no way to accurately assign a frequency to the different analyses each ambiguous form allowed. There was a relatively small number of morphologically ambiguous words in our sample (19 altogether), and an estimated 292 in the entire vocabulary of printed school English. Even if each of these were three ways ambiguous (definitely an overestimate), this would add less than a thousand items to the total population, and these would be scattered among various categories. Inclusion or exclusion of these items in our estimates therefore makes no meaningful difference in the size of the categories we will be considering.

[3]Main entries in Webster's Third meet the following criteria:

First, plurals and verb parts are included under the main entry of the uninflected word, unless they would fall alphabetically more than five inches away from the main entry, in which case they are listed as a separate main entry in their appropriate alphabetical order. For example, bows, although it is a regular plural of bow, is listed as a separate main entry, because there are more than five inches of intervening words, e.g. bowie, bower, bowel. The same principle is followed for comparatives and superlatives, as well as varients in spelling. This means that almost all irregular plurals or verb forms, as well as many regular plurals and verb forms, will be listed as separate main entries.

Homonyms are given separate main entries, distinguished by initial superscript numbers. However, to facilitate comparison with Dupuy's estimate of the number of main entries in Webster's Third, we will follow Dupuy in not counting homonyms as separate main entries.

There are two forms of run-on entries. First, idioms and phrases based on the main entry word are listed as run-on entries under that main entry. These phrases and idioms are given separate definitions. Second, certain derived forms are also listed under the main entry, namely, forms derived by suffixes such as -ness or -ly. Not all such derived forms are thus included under the main entry. For example, quickly is listed as a main entry separately from quick. The following criteria are used for including a derived form under the main entry as a run-on entry: First,

the derivatives have to occur in alphabetical order. This presumably means that a derivative which would be separated from its main entry by intervening words would have to be listed as a separate main entry. Secondly, such derivitives are given without definition, presumably because their meaning is totally predictable from the meanings of the base and the affix. Therefore, any derivative whose meaning was not thus totally predictable would be listed as a separate entry. This summarizes the principles according to which types are grouped into main entries or split into distinct entries.

As to the types of items included in the dictionary: First of all, only certain types of proper names are included. Names of persons and geographical place names are not listed in the dictionary. However, some other types of names are listed, for example, names of tribes and peoples, and words derived from names of persons or places. For example: The word witchita is included as a name of the Amerindian people, and as an adjective based on the city name, but the city name itself is not included as an item in the dictionary. The proper name Tito is not found in the dictionary, but the noun Titoism is. The name Tiv (a people in Africa) is included, as well as adjectives such as Wickliffian.

Arabic numerals are not included, with the following exception: certain compounds, for example, 3-D, are included, but alphabetized as if they were spelled out. Compounds such as ninety-one, ninety-two, and ninety-three are also included.

Symbols, combining forms (e.g., pseudo-) and symbols (as for elements) are also included as dictionary items.

Compounds are also given as separate main entries. This includes compounds which are written as two separate words, e.g., luna moth or heat exhaustion.

[4]In principle, Webster's Third inludes compounds containing numbers, alphabetized as if they were spelled out. In practice, there are very few such items in this dictionary, one example being 3-D. None of the items in our sample coded as "compounds containing numbers" would have been listed as entries in Webster's Third, so this entire category was excluded.

In the category of nonwords, 7 items in our sample were prefixes and suffixes that would be listed in Webster's Third. However, Dupuy's (1974) calculation of the number of main entries in Webster's Third, which we will be making use of, excludes such entries, so we will also exclude these from our estimate.

Only a very small fraction of the alternate spellings in our sample would have appeared as separate entries in Webster's Third. Most of them are either deliberate or accidental misspellings, or words spelled in some unusual way, for example with hyphens to show syllabification. The small percentage of items in the category of alternate spellings that would constitute separate dictionary entries was taken into account in our estimate of "Webster main entry equivalents."

Although Webster's Third does contain some words that might be considered "foreign," one criterion for coding an item in our sample as "foreign" was that it not be listed in Webster's Third. Therefore all items in this category are excluded from our estimate.

Regular inflections with distinct meanings, e.g., experienced, collected, heaping, conditioning, tried, are given separate entries in Webster's Third. Such items were therefore included in our count of "Webster main entry equivalents."

[5]There is some reason to believe that at least at the higher end of the scale, scores on Dupuy's test may underestimate an individual's true vocabulary size by less than a factor of seven. The single largest factor contributing to the difference between Dupuy's estimate of the number of words in English and ours was his exclusion of words that did not occur as main entries in all of the four large dictionaries he used. Presumably the words that were excluded on this principle would on the average be harder or less likely to be known than words which did appear as main entries in all four dictionaries. Therefore, Dupuy's sample of words would contain a higher proportion of easier words than would be drawn from a complete range of 88,500 word families.

On the other hand, as already mentioned, it is our estimate of the number of distinct word families that is about seven times greater than dupuy's estimate of the number of Basic Words in English. If one takes the position that distinct meanings should be counted as separate words, Dupuy's test underestimates the size of an individual's vocabulary to an even greater degree.

[6]Beck, Perfetti, and McKeown (in press) matched children from different intact classes on the basis of pretest scores. Some of the control subjects were drawn from a combined third and fourth grade class. This class may have had lower reading attainment than the other classes. It is well known that matching does not eliminate preexperimental differences when the populations sampled are different (cf. Campbell & Boruch, 1975).

[7]Anderson and Freebody (in press) have shown that good readers in the middle grades aggressively apply morphological principles to hypothecate meanings for unfamiliar words.

Table 1

A "Word Family" Found in Our Sample
(in alphabetical order)

add
ADD
add-oil
.
.
.
added
addend
addends
.
.
.
adding
Adding
.
.
.
addition
Addition
ADDITION
addition-subtraction
additional
additions
additive
additive-inverse
additives
Additives
.
.
.
adds

Table 2

Relationships Among Members of a Word Family

In Terms of Target Words and "Immediate Ancestors"

Target Word	Immediate Ancestor	Affix	Relationship
add	---	---	Morphologically basic word
Add	add	---	capitalization
add-oil	add	---	compound (first member)
add-oil	oil	---	compound (second member)
added	add	---	regular inflection
addend	add	end	suffixation
addends	addend	---	regular inflection
adding	add	---	regular inflection
Adding	adding	---	capitalization
addition	add	ition	suffixation
Addition	addition	---	capitalization
ADDITION	addition	---	capitalization
addition-subtraction	addition	---	compound (first member)
addition-subtraction	subtraction	---	compound (second member)
additional	addition	al	suffixation
additions	addition	---	regular inflection
additive	addition	ive	suffix replacement
additive-inverse	additive	---	compound (first member)
additive-inverse	inverse	---	compound (second member)
additives	additive	---	regular inflection
Additives	additives	---	capitalization
adds	add	---	regular inflection

Table 3

Categories of Relationships Among Words

Category	Examples	
	Target Word	Immediate Ancestor
Morphologically basic word	add	---
Simple capitalization	Think	think
Alternate spellings	cart-horse	carthorse
Alternate pronunciations	fishin'	fishing
Alternate form of word	soya	soy
Alternate form with s	towards	toward
Regular inflections	walks	walk
Irregular inflections	went	go
Regular comparatives & superlatives	taller	tall
Irregular comparatives & superlatives	best	good
Suffixation	frustration	frustrate
Prefixation	unknown	known
Compounds and contractions	farmhand can't	farm, hand can, not
Truncations	rhino	rhinoceros
Idiosyncratic morological relationships	prophesy	prophecy

Table 4

Analysis of the Word Frequency Book by Word-Relatedness Categories

Category	Sample N	Sample %	Corpus N	Population %	Population N
A. Categories that would be included in most definitions of "word."					
Morphologically basic	846	11.65	10,108	7.46	45,453
Idiosyncractic relation	72	1.00	860	1.01	6,167
Suffixation	722	9.94	8,626	7.62	46,431
Prefixation	233	3.21	2,784	4.01	24,457
Compounding & contractions	1,038	14.30	12,402	17.23	105,044
Truncations	16	0.22	191	0.19	1,144
Abbreviations	12	0.17	143	0.15	897
Subtotal	2,939	40.48	35,115	37.66	229,593
B. Categories that would have their own separate entries in most dictionaries.					
Irregular inflections	49	0.67	585	0.25	1,528
Irregular comparative & superlative	1	0.01	12	0.002	13
Alternate forms of words	8	0.11	96	0.18	1,072
Alternate forms with s	8	0.11	96	0.11	693
Semantically irregular pl.	8	0.11	96	0.02	136
"Scientific plurals"	2	0.03	24	0.02	145
Subtotal	76	1.05	907	0.59	3,587
C. Categories that would not normally occur as separate dictionary entries.					
Regular inflections	1,553	21.39	18,555	16.37	99,547
Regular comparative & superlative	46	0.63	550	0.51	3,149
Incorrect regular infl.	3	0.04	36	0.07	450
Simple capitalization	618	8.51	7,384	8.51	51,906
Alternate spellings	136	1.87	1,625	3.05	18,584
Alternate pronunciations	87	1.20	1,039	1.21	7,381
Subtotal	2,443	33.65	29,188	29.69	181,017

Table 4 (Cont'd)

Category	Sample N	Sample %	Corpus N	Population %	Population N
D. Categories relating to proper names					
Basic proper names	929	12.80	11,099	14.78	90,107
Derived proper names	88	1.21	1,051	1.18	7,215
Capitalizations homographic with p.n.'s	76	1.05	908	0.67	4,114
Inflectional and other varients of p.n.'s	302	4.16	3,608	4.74	28,869
Subtotal	1,395	19.21	16,667	21.38	130,305
E. Categories not normally counted as words					
Formulae & numbers	339	5.50	4,767	5.89	35,891
Compounds containing numbers	41	0.56	490	0.80	4,894
Nonwords	147	2.02	1,756	3.35	20,444
Foreign words	46	0.63	550	0.92	5,618
Subtotal	633	8.80	7,563	10.97	66,847
F. Miscellaneous categories					
Errors in WFB (duplicated entries)	6	0.08	6	---	---
Ambiguous words (excluding proper names)	19	0.26	227	0.05	292
Ambiguous proper names	2	0.03	24	0.004	27
Missing ancestores added	203	2.80	2,425	---	---
2nd meanings of ambiguous items added	51	0.70	609	---	---

Table 5

Derived Words Arranged by Relationship Category

and Degree of Semantic Relationships

	Relationship Categories				
	Suffix	Prefix	Compound	Idiosyncratic	Total
SEM 0	26,840	12,999	21,773	519	62,131
SEM 1	6,289	4,051	28,591	666	39,597
SEM 2	6,904	3,476	26,033	879	37,292
SEM 3	3,717	2,630	17,817	2,435	26,599
SEM 4	1,413	636	4,675	1,162	7,886
SEM 5	1,269	666	6,155	505	8,595
SEM 0-2	40,033	20,526	76,397	2,064	139,020
SEM 3-5	6,399	3,932	28,647	4,102	43,080

Table 6

Derived Words Arranged by Relationship Category

and Degree of Semantic Relationship

(Minimal Semantic Distance Based on Most Similar Meanings)

	Relationship Categories				
	Suffix	Prefix	Compound	Idiosyncratic	Total
SEM 0	28,491	13,555	22,436	807	65,289
SEM 1	6,780	4,296	32,132	627	43,835
SEM 2	6,562	3,523	25,223	1,178	36,486
SEM 3	2,646	1,828	16,387	2,774	23,635
SEM 4	740	456	2,765	673	4,634
SEM 5	64	13	2,820	65	2,962
SEM 0-5	41,833	21,374	79,791	2,612	145,610
SEM 3-5	3,450	2,297	21,972	3,512	31,231

Table 7

Some Estimates of the Number of Words in English

	Main Entries[a]	Basic Words[b]	Basic Words[c]	Total Words[c]	Basic Words[d]
Author's original estimate	240,000	12,300	166,247	370,265	99,600
Estimated number in the WFB	37,707	16,655[e]	31,095	50,765	18,037
Estimated number in printed school English	243,136	88,533[e]	192,909	344,572	91,466

[a]Webster's Third (estimated by Dupuy, 1974)

[b]Dupuy (1974)

[c]Seashore & Eckerson (1944)

[d]Seashore & Eckerson (1944) (with revision by Lorge & Chall, 1963)

[e]Morphologically basic words plus semantically opaque (SEM 3, 4, 5) derivaties

Table 8

Polysemy Among Morphologically Basic Words

Polysemy Measure	Extent of Polysemy			
	Mean Number of Meanings Per Morphologically Basic Word		Total Number of Distinct Meanings Among Morphologically Basic Words	
	WFB	Population	WFB	Population
SEM 0	4.218		42,636	
SEM 1	2.872		29,030	
SEM 2	2.038	1.615	20,600	73,417
SEM 3	1.417	1.316	14,323	59,821
SEM 4	1.231		12,443	
Homographs	1.103		11,149	
Phrasal and idiomatic entries	0.436		4,407	

Table 9

Count of Basic Words Incorporating Homophony

	Number of Words	
	WFB	Population
"Semantically distinct" defined with SEM 2 cut-off		
Number of distinct meanings of morphologically basic words	20,600	73,417
Number of distinct derived words	4,779	31,821
Total	25,379	105,238
"Semantically distinct" defined with SEM 3 cut-off		
Number of distinct meanings of morphologically basic words	14,323	59,821
Number of distinct derived words	1,039	7,596
Total	15,362	67,417

Table 10

Cumulative Distribution of Words by Frequency

Frequency (in terms of U)	Number of Words in Printed School English at or Above that Frequency		
	Graphically Distinct Types	Morphologically Basic Words and Semantically Opaque Derivatives	Semantically Transparent Derivatives
100.00	890	555	55
31.623	2,305	1,225	175
10.000	5,480	2,450	455
3.1623	11,980	4,330	1,290
1.0000	24,108	6,700	3,300
.31623	44,743	10,400	7,150
.10000	76,757	15,350	13,400
.03162	122,045	21,700	23,000
.00132	304,803	46,300	65,000
.00003	512,886	75,000	116,000
0.0000	609,606	88,500	139,000

Table 11

The Average Composition of a Word Family

Number of Words		Type of Words
Definition A	Definition B	
1.00	1.00	Base word (a morphologically basic word or semantically opaque derivative)
.15	---	SEM 4 derivatives
.49	---	SEM 3 derivatives
.65	---	Total semantically obscure derivatives (SEM 3, SEM 4)
.69	.42	SEM 2 derivatives
.73	.45	SEM 1 derivatives
1.15	.70	SEM 0 derivatives
2.57	1.57	Total semantically transparent derivatives (SEM 0-SEM 2)
.04	.02	Truncations and abbreviations
.07	.02	Irregular inflections, comparatives and superlatives; alternate forms of words; semantically irregular plurals
1.90	1.16	Regular inflections, comparatives and superlatives
2.00	1.22	Total inflections, abbreviations and truncations
.94	.58	Simple capitalizations
.34	.21	Alternate spellings
.14	.08	Alternate pronunciations
1.42	.87	Total minor variations in form
7.64	4.66	Total family size in graphically distinct types

Figure Caption

Figure 1. Graphic Representation of Relationships Among Words